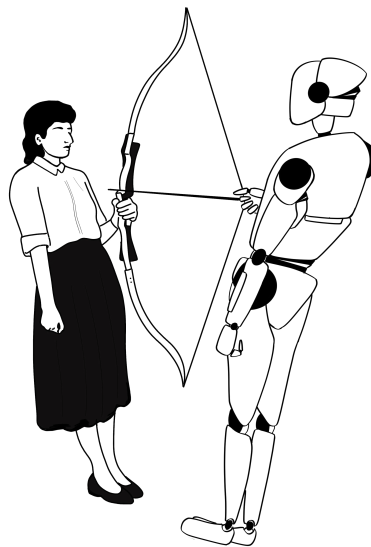


---

# *Trusting as a Moral Act: Trustworthy AI and Responsibility*



**Sara Blanco**

Dissertation  
zur Erlangung des akademischen Grades  
Doktor der Philosophie  
in der Philosophischen Fakultät  
der Eberhard Karls Universität Tübingen

---

**Advisors:** Prof. Dr. Wolfgang Spohn  
PD Dr. Eric Raidl

---

27th September 2024

Artwork: **Sara Blanco**

Cover art inspired by Marina Abramović and Ulay's performance *Rest Energy* (1980).

# **Trusting as a Moral Act: Trustworthy AI and Responsibility**

**D i s s e r t a t i o n**

**zur**

**Erlangung des akademischen Grades**

**Doktor der Philosophie**

**in der Philosophischen Fakultät**

**der Eberhard Karls Universität Tübingen**

**Vorgelegt von**

**Sara Blanco**

**aus**

**Valladolid, Spanien**

**2025**

**Gedruckt mit Genehmigung der Philosophischen Fakultät  
der Eberhard Karls Universität Tübingen**

**Dekanin: Prof. Dr. Angelika Zirker**

**Hauptberichterstatter: Prof. Dr. Wolfgang Spohn**

**Mitberichterstatter: PD Dr. Eric Raidl  
Prof. Dr. Hong Yu Wong**

**Tag der mündlichen Prüfung: 26.02.2025**

**Tübingen,**

**Universitätsbibliothek: TOBIAS-lib**

---

# Contents

---

<b>Introduction</b>	<b>3</b>
<b>1 Trust and related concepts</b>	<b>9</b>
1.1 Introduction	11
1.2 Reliance	12
1.3 Trust	14
1.4 Trustworthiness	16
1.5 The relationship between trust and trustworthiness	17
1.6 Trust versus Reliance	18
<b>2 Kinds of trust</b>	<b>21</b>
2.1 Introduction	23
2.2 AI as particular systems	24
2.3 AI as a socio-technical tool	27
2.3.1 Users, tools and systems	27
2.3.2 Quasi-agents	29
2.4 Trust in AI	31
2.4.1 Typical scenarios involving artificial trustees	31
2.4.2 A tentative description of "trust in AI"	32
2.4.3 Trust in AI and trust in humans: shared features	34
2.5 Interpersonal Trust	36
2.5.1 Trust as risk-assessment	37
2.5.2 Trust as an attitude	40
2.5.3 Trust as a contract	42
2.5.4 Recapitulation	45
2.6 Institutional Trust	46
2.7 Revisiting trust in AI	49
2.7.1 Motives	50
2.7.2 Agency	51
2.7.3 Rational reasons	54
<b>3 Trust as a belief-based disposition</b>	<b>57</b>
3.1 Introduction	59
3.2 Motivation	62
3.2.1 Challenges	62
3.2.2 Tensions to overcome	63
3.2.3 Resolution: a doxastic account	66
3.3 Proposal: a belief-based disposition	68
3.3.1 Preliminary problems	68

3.3.2	Trust as a belief-based disposition	68
3.3.3	Challenges check	77
3.4	Possible objections	78
<b>4</b>	<b>Building trust</b>	<b>83</b>
4.1	Introduction	85
4.2	The value of trust	87
4.3	The Explainability-Trust Hypothesis	88
4.3.1	Background and formulation	88
4.3.2	Epistemic limitations	90
4.3.3	A new version of ET	94
4.4	Degrees of trust as degrees of trust belief	97
4.5	The value of distrust	99
<b>5</b>	<b>Alternatives to trust</b>	<b>103</b>
5.1	Introduction	104
5.2	Nickel, Franssen and Kroes: a thin notion of trust	105
5.3	Ryan: rational trust	110
5.4	Durán and Jongsma: computational reliabilism	114
5.5	Recapitulation	117
<b>6</b>	<b>The moral component of trust</b>	<b>119</b>
6.1	Introduction	120
6.2	Trusting relationships and moral responsibility	122
6.2.1	Moral responsibility as a relational concept	122
6.2.2	Trust as normative expectation	123
6.2.3	Trust as an invitation	125
6.3	Trust in AI as a moral concept	126
6.3.1	Portraying trustworthiness	126
6.3.2	The moral significance of portraying of trustworthiness	128
6.4	Artificial trustees and responsibility	131
6.5	Revisiting the BBD account	135
	<b>Conclusion</b>	<b>137</b>
<b>A</b>	<b>Trust: Related Concepts</b>	<b>141</b>
A.1	Distrust	142
A.2	Mistrust	142
A.3	Failed trust	143
A.4	Misplaced trust	143
<b>B</b>	<b>Dispositions and Beliefs</b>	<b>145</b>
B.1	Dipositions	146
B.1.1	Brief history of dispositions	146
B.1.2	Single versus multi-track dispositions	147
B.2	Beliefs	148
B.2.1	Beliefs as multitrack dispositions	148

<b>B.2.2 Dispositional beliefs</b> . . . . .	148
<b>Glossary</b>	151
<b>Bibliography</b>	153



---

## Acknowledgments

This thesis is the tangible product of the last four years of my life. I had the chance to spend those years doing what I enjoy the most: learning. For this opportunity, I would like to thank the University of Tübingen, the Cluster of Excellence ‘Machine Learning—New Perspectives for Science’ and the Baden-Württemberg Stiftung for their financial support, and for providing me with the opportunity to pursue this project. I could have never finished this thesis without the material means to do so, and for that, I’m deeply grateful. I’m grateful for the opportunity to dedicate this time to learning full-time about a topic that I’m passionate about and that has allowed me to grow as a researcher and as a person. I’m very well aware that not everyone is this lucky, and that there are many places in the world where such opportunities are rare and reserved for students much more exceptional than myself. I am also deeply grateful to Wolfgang Spohn, for taking me under his supervision at a very challenging time. His guidance and encouragement were invaluable for me to continue this path. I owe a special thanks to Eric Raidl for his unwavering support throughout this process, and for his infinite patience and kindness in reviewing my work, always helping me to improve and refine my ideas. I would also like to extend my sincere thanks to Hong Yu Wong, who not only stepped in during the final year of my PhD but also went above and beyond in offering guidance and support. His dedication made a lasting difference, and I am truly grateful for his help during these critical months and for encouraging me to hang in there. I would also like to acknowledge the members of the Ethics and Philosophy Lab and the AITE project for their input and professional support throughout this process.

Last but not least, I would like to thank the friends who put up with my four years of whining and constant self-doubt, especially Nina, Niki, Pablo and Maurice. A special thanks to Cristina, my therapist, whose help was invaluable in getting me this far. I am equally grateful to my mother, who witnessed my struggles, excitement, and endless talk about topics she may not have cared much for, but listened to because she cares so deeply about me. In a way, I’m also grateful to my father, who didn’t see me start this PhD and won’t see me finish it, but who always believed that one day I would make it.

During the completion of this thesis, I was funded by the Baden-Württemberg Foundation (program “Verantwortliche Künstliche Intelligenz”) as part of the project AITE (Artificial Intelligence, Trustworthiness and Explainability). I was also supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence ‘Machine Learning—New Perspectives for Science’, EXC 2064, project number 390727645).

---

## *Introduction*

---

In recent years, artificial intelligence (AI) has become an integral part of our daily lives. Today, one can simply ask their phone to make an appointment, and an AI-powered assistant, such as Google's Duplex (Leviathan and Matias [2018](#)) or Gemini (Kattukaran [2024](#)), will execute this task seamlessly. This system is capable of processing the user's request, extracting necessary data from the internet, and using this information to place a call. During the call, the system interacts with the human respondent, conveying the required information accurately. The possibility to delegate such complex tasks to AI holds the promise of enhancing our efficiency and simplifying our lives. However, no reward comes without risk. In the case of AI, some of the main risks include compromising privacy, inaccuracy, and bias, among others. This raises a crucial question: can (and should) AI be trusted? This twofold question addresses two distinct issues that require a two-step answer:

- (1) Is it possible to apply the concept of trust to AI systems?
- (2) Which AI systems should we trust and on which basis?

To address the second issue, it is necessary to clarify what it means to trust AI. When exploring this question, it becomes necessary to address the moral component of trust. In trusting relationships, the trustor makes themselves vulnerable to the trustee because the trustor depends on the trustee in some domain. This setup implies that the trustee has the moral responsibility not to exploit the trustor's vulnerability. This moral component also applies in the case of trust in AI. I will argue that it is a key aspect of trust that makes the concept suitable to be extended to human-AI interaction. In this thesis, I intend to explore the above questions: both about the meaning of trust in AI and the role of moral responsibility in human-AI relationships. To achieve this, I will proceed in the following manner.

In chapter [1](#), I introduce three central concepts: reliance, trust, and trustworthiness. I do so because, to investigate trust in AI, it is imperative to first clarify the concept of trust more broadly. To this end, it is also necessary to define the related concepts of reliance and trustworthiness. Initially, I will offer a tentative definition of trust as a type of relationship. Trusting relationships are characterised by the trustor believing that they could delegate some task to the trustee and, under certain circumstances, being willing to do so. Other accounts define trust as an attitude held by the trustor, to which the trustee is relatively alien (Baier [1986](#); Jones [1996](#)). In contrast, accounts that see trust as relational often describe it as a form of cooperation (Gambetta [1988](#); Hardin [2002](#)). Unlike these views, I will focus on the relationship between the trustor and the trustee, broadening the kind of circumstances in which such a relationship can occur. The concepts of trust, trustworthiness and reliance are pivotal throughout the thesis, recurring frequently and undergoing continuous

discussion and redefinition. In this first chapter, I will provide intuitive definitions and a brief overview of how the concepts relate to each other. In particular, I outline the relationship between trust and trustworthiness and the distinction between trust and reliance. My approach to trustworthiness is novel, as I consider it subordinate to trust rather than the other way around (McLeod 2021). This chapter will also lay the foundation for a more in-depth discussion on trust in AI, allowing me to contrast trust and reliance in the context of AI systems.

In chapter 2, I will continue laying the basis for the rest of the thesis, this time by introducing the concepts of AI and trust in AI. Regarding AI, I will clarify my use of the term by making the distinction between *AI as particular systems* and *AI as a socio-technical tool*. When I discuss AI as particular systems, I refer to specific models –such as chatGPT– rather than AI as a kind of model. More specifically, my focus is on machine learning (ML) algorithms and, in particular, on neural networks (NNs). When I discuss AI as a socio-technical tool, I am referring to an abstract concept. I will use this terminology to refer to the hybrid nature of AI systems, which integrate both technical and social components. I use the term *socio-technical tool*, instead of *socio-technical system*, to distinguish AI from previous technologies that already incorporated a social factor (see Van House (2003, p. 72)). What makes AI special is that it constitutes a tool that is able to improve itself and ‘learn’ from social inputs. In this way, AI can use data to achieve results that are unpredictable even for its designers. Some authors consider this perceived autonomy as some sort of derived<sup>1</sup> agency or *quasi-agency*. According to Coeckelbergh (2009), humans treat non-humans *as if* they were agents in order “to support the (quasi)moral and social dimension of our dealings with them” (p. 184). By ‘quasi-agency’, I refer to this kind of ‘as if’ approach. If humans interact with AI systems ‘as if’ they made decisions, it raises the question of whether AI systems can be objects of trust in a similar way to human agents. I will use two case studies to posit the concept of ‘trust in AI’<sup>2</sup> and to outline the main features to be expected from trust relationships between humans and AI (given that these relationships can be indeed trusting ones). I will argue that trust, both in humans and in AI, is characterised by six key features: the trustor’s positive expectations, the (potential) wish of delegation, the risk of the trustee failing to the trustor, being motives-based, being grounded in rational reasons and intersubjectivity. The last three points may be topic of debate, so I argue for them in chapter 2. In particular, I will pay special attention to the notion of *motives*, as my use of the term differs from the more widespread use of this term in the trust literature. In this second chapter, I will offer my first characterisation of ‘trust in AI’. To do so, I need to first discuss the main accounts of both interpersonal and institutional trust. Then, I will compare the main characteristics of these concepts with the concept of ‘trust in AI’, looking for both similarities and differences.

In chapter 3, I develop my own account of trust. There is a gap in the literature of trust concerning non-human trustees. Consequently, when the public and parts of academia began discussing human-AI interactions in terms of trust, there has been a

---

<sup>1</sup>Derived from the humans who are (partially) responsible for the AI’s outcomes, such as its designers.

<sup>2</sup>In this context, I will use ‘trust in AI’ in single quotation marks until I present a solid argument for using trust as a legitimate concept to describe human-AI relationships.

tendency to adapt theories of interpersonal or institutional trust to the discourse on AI. However, AI systems constitute a different kind of trustee. To understand trust in AI as a distinct form of trust, I take a step back and propose a broader account, characterising trust as a *belief-based disposition* (BBD). Belief-based dispositions are the kind of dispositions that originate in belief. That means that belief-based dispositions are states in which an object  $x$  is disposed to  $M$  when  $C$  only under the background condition of a certain belief being held. Applied to trust, this means that the trustor trusts the trustee when the trustor is disposed to delegate some task  $\delta$  to the trustee (when the trustor needs or wishes to do so), if the trustor holds a certain belief. In the case of trust, such a belief would be a *trust belief*. A trust belief is the trustor’s belief that the trustee is trustworthy. The trustor considers the trustee to be trustworthy if the trustee is perceived as capable of performing  $\delta$  and their motives to do  $\delta$  are approved by the trustor. I speak of ‘perceived’ capabilities and motives because the trustor has no direct access to the internal states of the trustee and can only infer them based on their own perception. This means that, ultimately, the trustor is never certain whether the trustee is trustworthy or not, but they can have access to reasons that allow them to infer the trustee’s (un)trustworthiness. This detail is relevant because it allows me to introduce the distinction between trust (in a general sense) and *appropriate trust*. In the BBD account, I consider trust to be appropriate when the trustor’s trust belief is justified. That is, the trustor has access to reasons that justify believing that the trustee is trustworthy. This is an important nuance regardless of the nature of the trustee, which could be a person, an institution or an AI system. The goal of the BBD account is to describe trust both in human and AI trustees. In this chapter, I approach trust in a more holistic way, and then zoom in on the particular case of trust in AI in the following chapters.

In chapter 4, I explore the literature on trust in AI to identify methods for fostering trust in AI systems. While in the previous chapter I addressed the question of what trust is, in this chapter I aim to answer how trust can be built, particularly in the case of artificial trustees. Designing trustworthy AI systems is challenging, especially due to the black box problem. The black box problem concerns systems whose inner workings are not understandable by humans. This opacity means that even highly accurate systems may still not be fully trusted. Recently, explainable AI (XAI) has emerged as a promising approach to address this issue and enhance trust in AI systems. The underlying assumption behind the promise of XAI systems is that providing explanations can help cultivate trust. I refer to the latter assumption as the *Explainability-Trust Hypothesis* (ET). ET can be defined as the assumption that explainability is a suitable means for facilitating trust in a potential trustor. I will uncover two epistemic limitations of ET:

- (a) Only explanations tailored to the epistemic background of the trustor have the potential to generate a trust belief.
- (b) Only *good*<sup>3</sup> explanations have the potential to generate a trust belief in the explanation recipient.

---

<sup>3</sup>What is a good explanation constitutes a rich debate in epistemology. The details of this debate go beyond the scope of this thesis. However, in chapter 4, I will attempt to at least clarify what I am referring to by *good explanations* in the context of XAI.

I will also provide further details on how trust beliefs are built and, specifically, how different degrees of trust correspond to different degrees of trust belief. By examining the nuances of trust belief formation, we can understand how varying trust degrees relate to the beliefs that lead the trustor to perceive the trustee as trustworthy. It is important to note that, depending on the basis of the trust belief, trust will be more or less justified. This deeper understanding of trust beliefs supports my notion of *appropriate trust*. Building on my ideas on how trust beliefs are formed and how they should be formed, in this chapter, I insist on the idea the trust placed in AI systems should be warranted. That is, when it comes to AI trustees, the goal should not be mere trust but *appropriate trust*.

In chapter 5, I will set the stage for the final claim of this thesis: trust has a moral component that makes it a suitable concept to describe human-AI relationships. Before arguing for that idea, in chapter 5, I will address some influential arguments against it. I will discuss the work of authors who claim that trusting AI (or technology) is not possible but, at best, AI can only be relied on. For this purpose, I have chosen three papers by Nickel, Franssen, and Kroes (2010), Ryan (2020), and Durán and Jongsma (2021). I analyse the arguments of these authors and explain my disagreement, going back and forth between my BBD account of trust and the characterization of trust, reliance and motives presented in chapters 1 and chapter 2. I will argue that reliance, as I presented it at the beginning of this thesis, does not suffice to describe the kind of relationships that would be desirable to have in some human-AI interactions. Furthermore, most of the points raised by these authors do not pose a problem if trust is understood as I proposed in chapter 3.

In chapter 6, I explore the moral component of trust, particularly of trust in AI. I will argue that this moral component makes trust a suitable goal for human-AI relationships, unlike reliance.<sup>4</sup> I will reexamine the concept of trust in more detail, this time paying special attention to the relationship between trustor and trustee and the moral responsibility that the latter has towards the former. Instead of focusing on the consequences of actions, I will concentrate on the relationships between the parties involved to ascribe responsibility. This approach is particularly suitable when one party is an AI, aligning with my framing of trust in AI within the context of human-AI interactions. In this final chapter, I will examine the normative expectations inherent in trusting relationships and revisit the concept of trust as an invitation that I introduced in chapter 2. My focus is on the trustee's role in shaping the trustor's beliefs. The main idea is that the trustee has the moral responsibility to accurately represent their trustworthiness to the trustor. That is, the trustee is responsible for how they present themselves to the trustor. Portraying an accurate image of one's trustworthiness is morally significant<sup>5</sup> because it triggers others to act in consequence. In other words, the trustee should not mislead the trustor about their trustworthiness because it would cause the trustor to act based on false information, leading to decisions they would not make if they knew the true facts. Based on this

---

<sup>4</sup>In some high-stake scenarios. In certain cases, reliance may be sufficient, but my main point is that reliance is not necessarily the optimal scenario for human trustors.

<sup>5</sup>Throughout this thesis, I follow the use of 'moral significance' as employed by Persson (2021). Further details will be provided where relevant.

argument, I consider trust to be a moral concept since, in trusting relationships, the trustee has a moral responsibility towards the trustor. I will explore this kind of moral responsibility in the particular case of artificial trustees. Additionally, I will revisit the idea of AI as a socio-technical tool. I will provide further details on my understanding of artificial trustees, aiming to clarify how responsibility attribution works in these cases. Finally, I integrate the moral aspects of trust into my account of trust as a belief-based disposition (BBD), aiming to provide a conceptual framework that helps to understand trust in AI, together with its moral implications.

In sum, the main aim of this thesis is to provide a conceptual framework for trust in AI that clarifies the nature of human-AI trusting relationships and highlights the role of moral responsibility within them. To achieve this, I propose a novel account of trust that explains not only cases of human-AI interaction but also the more paradigmatic instances of interpersonal and institutional trust. My BBD account of trust aspires to be a holistic theory that sheds light on the variety of questions arising from trusting relationships, regardless of the trustee's nature. These questions involve complex philosophical issues, such as the concepts of motives and agency. By adopting a broader understanding of these concepts, I aim to preserve well-established conclusions from the trust literature while adapting its key elements to my theory. This new perspective on widely debated concepts will allow me to introduce fresh discussions into the literature on trust. An important discussion in this regard is how trust is built. While there is no consensus on how to build interpersonal and institutional trust, incorporating AI into the conversation further complicates the matter. I hope that the BBD account can lay the foundation for clarifying both longstanding and emerging issues in trust, and contribute to the field of XAI by addressing overlooked aspects of the relationship between trust and explanations. Conceptualising AI as a socio-technical tool is a crucial part of this endeavour, particularly when examining the moral dimension that arises from the inherent vulnerability in trusting relationships. Although I argue that trust is a suitable concept for describing human-AI relationships, the moral responsibility that trust often entails can only be traced back to part of the artificial trustee. Specifically, I will show that this part refers to the social component of the socio-technical tool: the human agents responsible for the AI system's functioning.



# 1

---

## *Trust and related concepts*

---



---

## 1.1 Introduction

Trust is often considered the glue of a healthy society. We need others and others need us. People often encounter tasks that they cannot fulfil by themselves, either because of a lack of skills, knowledge or time. It is in such situations that the need for delegation arises. Trust is often paired with delegation. People are willing to delegate tasks to those who they trust. Usually, this will to delegate rests on the belief that the trusted person is capable and motivated to perform the task they are trusted with. In this chapter, I will argue that trust is paired with the possibility of delegation and not necessarily with actual delegation. Generally, trusting relationships involve two people. It is unclear whether the concept of trust can suitably be extended to describe relationships between humans and other entities, such as artificial intelligent systems. However, both the general public and a part of academia often talk about trust and distrust in artificial intelligence (AI).

AI constitutes more than a mere technical artefact. Due to its potential to affect people's lives, AI is considered by many a socio-technical tool (HLEG 2019). This means that its successful implementation concerns the interaction of not only technical, but also social factors, and interactions between these. From this perspective, it is crucial for the successful implementation of AI to conceptualise the kind of relationship that we, as a society, aim to have with these kinds of systems. During the last decades, the concept of trust has gained considerable attention in the AI community. Numerous ethic guidelines, white papers and scientific publications<sup>1</sup> highlight the importance of developing trusting relationships with non-human entities for a successful implementation of AI in society. However, works coming from the AI field rarely take into account the literature about more traditional notions of trust, which is rich and extensive. Conceptually, trust has been approached by different disciplines, such as sociology, political science and philosophy. Most of what is written about this topic concern cases of trust among humans. But when it comes to AI, it remains unclear how to conceive trust. Thus, trust in AI appears to be the cornerstone of ethical debates on AI, without a clear grasp of what is exactly being referred to.

I will show that it is hard to fit the possibility of trusting non-human entities into the existing traditional frameworks of trust. The reason is that most of these frameworks have been developed around cases of trust in persons and institutions. Without a proper theory of trust that reflects cases in which trust is placed on non-humans, it remains unclear what it means to trust AI. With the lack of consensus on the latter, trust becomes an empty goal to aim for in AI. The question, therefore, arises whether trust is a suitable term to describe the normative relationships that humans should have with artificial intelligent systems. In other words, what kind of relationships humans should have with AI, attending to ethical and social factors.

In this chapter, I introduce the concepts that will allow me to discuss this central question: reliance, trust and trustworthiness. The chapter is structured as follows.

---

<sup>1</sup>E.g. OECD (2017), HLEG (2019), Leslie (2019), United Nations (2021), and Gunning et al. (2019)

In section 1.2, I introduce the concept of reliance. I will discuss the main features of reliance and the kind of situations in which it occurs. In section 1.3, I introduce the concept of trust and offer a tentative definition. In this first characterization, I will connect the notion of trust to closely related concepts, such as delegation, risk or vulnerability. In section 1.4, I look in detail at the idea of trustworthiness. I will define the concept, in line with my previous definition of trust. In section 1.5, I explore the relationship between trust and trustworthiness and their normative intertwining. Finally, in section 1.6, I distinguish between trust and reliance, which is key to understanding why relying on AI does not equal trusting it. On this basis, in chapter 2 I will come back to the question of whether trust is a suitable concept to describe human-AI relationships.

---

## 1.2 Reliance

In our daily life, we constantly engage in activities that require the use of something or someone else. The more complex or ambitious a goal is, the more resources we need to achieve it. Often, we become dependent on such resources since, without them, it would not be possible to achieve what we want. I cannot reach my lamp without climbing a ladder and I cannot learn German without the guidance of someone who already speaks it. This state of dependence on something or someone else is called reliance.

Intuitively, a person relies on another one when the first one needs to delegate something to the second one.<sup>2</sup> That is, I rely on you if I need you to do something for me. Reliance can be applied both to objects and humans (Hawley 2014; Nickel, Franssen, and Kroes 2010). In natural language, we use the verb ‘to rely on’ both for cases in which we depend on someone else to do something, and cases in which we need an object to achieve something.

Reliance is often associated with regularity and expectations. When I rely on someone, I expect them to act according to my plans (Holton 1994, p. 65). Generally, expectations arise from the observation of regular behaviour. It is natural to expect reality to continue being the way it was observed, even though reality may always continue in another way. Because the sun came up every morning, I expect it will come up tomorrow. In the same way, I expect my ladder to hold my weight as it has always done or my German teacher to continue telling me true facts about the German language. We rely on others when we expect them to behave as we want, regardless of what motivates them to do so.

Reliance is based on expectations, and expectations do not always match reality. It is possible to rely on a person who ends up not fulfilling our expectations, even though we believed they will. When we realise we have relied on someone we should not have, we may feel let down. However, as Hawley (2014, p. 2) points out, when reliance is broken, we have no right in feeling betrayed. Let me illustrate this with

---

<sup>2</sup>In every case of reliance, there is a need (or desire) for delegation. But the need or desire to delegate does not always give rise to actual reliance. I may learn how to do the task myself, or simply chose not to delegate, even though I need to.

an example.<sup>3</sup> Allegedly, Immanuel Kant was an extremely punctual person. Every day of his life, he passed by the Lindenallee in his natal Königsberg exactly at the same time. It is commonly told that Kant's neighbours relied on this daily habit of his to adjust their clocks. But imagine that one day, Kant overslept and passed later than usual by the Lindenallee. As a consequence, the reliance his neighbours placed on him had been broken. In this scenario, it is hard to picture Kant's neighbours resenting Kant, let alone feeling betrayed by him.

Usually, we rely on others because we need to. This means that reliance is an action of the will, that is, it can be chosen, or not (Baier 1986, p. 234). Sometimes, people rely on something or someone simply because they have no alternative. If I need to change the bulb of my lamp and the only ladder I have is old and shaky, I may not feel confident about using it. Even so, if that ladder is my only means to reach the lamp and I really need to have some light, I may resign myself and rely on that ladder, even though I thereby take the risk of falling.

Reliance is also involved in a wide spectrum of situations. Depending on who we rely on and for what, our reliance may be more or less confident. Sometimes we rely on others almost automatically, not even thinking about it. But on occasions, thinking about who to rely on is crucial. Using the concept of reliance when describing human-AI relationships is relatively unproblematic. Nowadays, we rely on AI as we previously relied on other types of technology or tools. In line with the intuitive characterization of reliance that I offered so far, relying on AI means depending on it to achieve a goal. When a user relies on an artificial system, they delegate a task to the system to achieve results that they could not achieve by themselves. In order to capture the concept of reliance both for human-human and human-object (including AIs) interaction, I propose the following general definition of reliance:

**Definition 1.1.** *Reliance* is the dependence that one party has towards another when the first needs to delegate something to the second.

According to the definition above, reliance only happens when there is something that needs to be delegated. Here, I understand delegation in a broad sense since, strictly speaking, one cannot delegate something to inanimate objects. However, we can depend on objects to achieve something and, in this sense, we 'delegate' part of the process to achieve our goal to the object we are using. For example, imagine that my goal is to change a light bulb. In order to do so, first I need to reach a certain altitude, then I need to grab my lamp and finally, I need to unscrew the old bulb and replace it with a new one. I am a rather short person, hence I cannot reach the altitude I need on my own. Therefore, I 'delegate' the task of supporting my weight at the required altitude to a ladder, so I can reach my lamp and change the bulb. In this sense, I rely on the ladder.

---

<sup>3</sup>Example taken from Baier (1986, p. 235).

---

## 1.3 Trust

Close to the concept of reliance, we find the notion of trust. Most authors distinguish trust from reliance.<sup>4</sup> In this section, I proceed to describe what I understand by trust.

As already mentioned, no one is able to do by themselves everything they would like to do. Many of our goals exceed our own capacities and, therefore, we need the help of others. Ideally, one asks for help from those who are able and willing to offer that help. And often, one does not even need to ask: there are people who relate to us in such a way, that we are confident that they would do something for us if we need them. When this is the case, a trusting relationship exists.

Trust is the kind of relationship that one person has with another one when the former can count on the other and is willing to rely on the latter (if necessary). The will to rely on someone else comes with the assumption that the other person will do something for the person who trusts. The one who trusts is called the *trustor* and the one who is trusted is called the *trustee*. The trustor thinks that the trustee would do something for them because the trustor considers that the trustee is capable to do that something. Moreover, the trustor thinks that the trustee has the right motivation to do so, so their interests are aligned in some relevant aspect. For example, imagine that I trust my neighbour to take care of my cat while I am away. I trust my neighbour because I consider they have the necessary skills to properly take care of my cat, and because I think that, as myself, my neighbour cares about the wellbeing of my cat. I believe they will do their best in my absence, therefore I trust them.

We often need to delegate endeavours to others. That is, others do things for us that either we cannot do by ourselves, or it is just convenient to use someone else's help. Ideally, we delegate our endeavours to those who we trust. Delegating something to someone means asking them to commit to that something. We expect others to do what we ask them to do. Then, ensuing our expectations, we rely on others more or less confidently. However, trust seems to go beyond mere reliance. Unlike in reliance cases, in trusting relationships, delegation does not always materialise. A trustor can trust a trustee but never need something from them, so there is no need to delegate anything. Maybe I will never be away, so I will never need my neighbour to take care of my cat. But as the trustor, I expect that if I ever need to, I can rely on my neighbour –the trustee. Thus, trusting relationships are not founded on actual delegation, but on the possibility of delegation.

In the cases in which the trustor trusts a trustee to whom no task is being delegated, what does their trust mean? It means that the trustor finds it reasonable to hypothetically delegate some task to the trustee. How does the trustor know whether is reasonable or not to delegate something in a hypothetical situation? Because of the motives that they attribute to the trustee in the current situation. This is why I consider trust to be motives-based. Trust refers to hypothetical situations in which trustors can only speculate about others' behaviours based on the available informa-

---

<sup>4</sup>See Baier (1986), Jones (1996), Holton (1994), O'Neill (2002), Hieronymi (2008), and McLeod (2021).

tion, namely, the trustee's current motives (deduced from their current behaviour or testimony).

In trusting relationships, the expectations that the trustor has of the trustee make the trustor being confident about their reliance. Usually, the past experiences that the trustor has with the trustee make them develop certain expectations. If my neighbour took care of my cat before, I expect them to be willing to do so again if I asked them. So there are reasons for the trustor to trust, and since they often refer to their own experience and judgements, such reasons are subjective. My trust in my neighbour is not universal; even if I tell another person how good with cats my neighbour is, they may still not trust my neighbour as I do because they do not have the first-person experience that I have. My neighbour and I have a particular relationship that no one else has. Then, no one believes what I believe about my neighbour because of the same reasons that I do. Other people may agree that my neighbour is good with cats, but maybe they do so because my neighbour also took care of their cats at some point, or just because they believe what I say about such neighbour (so they trust my neighbour because they trust me). Thus, trust is not only about the trustor delegating something and the trustee potentially doing that something. The reasons to be *willing* to delegate play a key role in trust.

Furthermore, trust happens in the context of uncertainty. If I know how someone will behave, I do not need to trust them. Trust is always related to future hypothetical scenarios. I trust my neighbour because of how I expect them to behave in case I ever needed their help. The trustor has expectations about the trustee's behaviour under certain circumstances. But expectations do not always match reality. The hypothetical scenario could become actual in the future. And still, the trustor's hypothetical behaviour that the trustor expected may become a reality or not. When a trustor trusts a trustee to do something, there is always the risk that the trustee fails to do that something. The trustor is aware of the risk, but they still trust because they believe that it is unlikely that the trustee does not act as expected.

By trusting, the trustor accepts the risk that the trustee may not behave as the trustor expects. Accepting this risk means willingly making themselves vulnerable towards the trustee. For this reason, Annette Baier (1986, p. 235) describes trust as an accepted vulnerability. As I said, trust refers to uncertain scenarios, to expectations that may materialise or not. In the end, the trustor does not know what the trustee will do. When I trust someone, I give them the power to betray me. I expect them not to betray me, but I am aware that they could. Betrayal could take place if I am mistaken about the trustee's intention or motivation. I thought their interests matched mine, but in reality, they did not. Maybe I thought higher of my neighbour than I should, but they actually do not care about my cat enough to feed it every day. The trustor can also be mistaken about the trustee's capacities. Maybe the trustee wants in fact to comply with the trustor's desires, but they simply cannot, while the trustor believed otherwise. My neighbour may care about my cat, but I may be unaware of how busy their life is; too busy to commit to the responsibility of looking after my cat.

Based on the discussion above, I propose the following working definition of trust:

**Definition 1.2.** *Trust* is the kind of relationship that a trustor has towards a trustee

when the trustor believes that they could delegate something to the trustee and they would be willing to do so if necessary.<sup>5</sup>

The definition above does not go into too much detail about the nature of trusting relationships, but it suffices for the argumentative purposes of this chapter.

---

## 1.4 Trustworthiness

We have seen that trust happens in contexts of uncertainty. A trustor trusts a trustee if the trustor believes that the trustee will behave as the trustor wants, even though the trustor does not know whether that will be the case. In other words, trustors trust trustees who they perceive as trustworthy. I proceed now to develop what I understand by trustworthiness.

Trustworthiness is the relational property of being worthy of trust. A trustee is worthy of a trustor's trust if they behave as the trustor expects, motivated by reasons that the trustor considers good. Going back to the example with my neighbour, if they take care of my cat –motivated by their sense of responsibility and love for cats– I would perceive my neighbour as trustworthy.

Being trustworthy and appearing as such are not the same thing. Someone can seem trustworthy without actually being trustworthy and vice versa. We tend to trust people who we perceive as trustworthy, and to generally distrust those who we do not. In her BBC lectures, O'Neill (2002) rightly points out that distrust does not necessarily come from untrustworthiness, but from a whole context of suspicion. According to her, at the time, society was going through a “supposed crisis of trust”. The root of the crisis was not the decrease in the trustworthiness of the media (and the politicians, institutions, etc.), as is commonly argued. There are trustworthy and untrustworthy trustees, as there have always been. What has changed is the difficulty to discern the former from the latter. Twenty years later, this difficulty is even higher with the appearance of fake news and deep fakes. Being able to distinguish trustworthy trustees from untrustworthy ones is a hard endeavour, and it is key to placing trust correctly.

The main difficulty when discerning trustworthiness is that trustworthiness is intertwined with trust and, as I showed in the last section, trust refers to future hypothetical scenarios. Being trustworthy is being worthy of trust. A trustee is worthy of trust if they *would* behave as the trustor wants to, in case hypothetical circumstances become actual. But I do not know what others would do in hypothetical scenarios, I can only speculate about the behaviour I expect from them. I think that, in the hypothetical case I needed my neighbour to take care of my cat, they would do so. I expect a certain behaviour from them, but I cannot foresee the future, so I do not know what would actually happen. I think my neighbour is trustworthy, but I cannot really prove that they are. Usually, the trustor has information about

---

<sup>5</sup>From now on, I will often use the phrasings such as ‘delegation materializing’, ‘delegation taking place’, ‘leading to delegation’ or similar, in which the term ‘delegation’ is used as an abbreviation of ‘the fact that trustor delegates something to the trustee’.

the trustee, such as past experiences between the two that make the trustor see the trustee as trustworthy. If the trustee really is like the trustor perceives them (really capable, really motivated), then the trustee does not only appear trustworthy but they actually are. Trustors make hypotheses about trustees. Whether those hypotheses are true or not cannot be known until the hypothetical scenario actualizes and the trustor has the opportunity to witness the actual behaviour of the trustee. However, even if the trustors cannot be sure about the trustees' trustworthiness, the belief that a trustee is trustworthy can be more or less reasonable. By reasonable, I mean that there are strong arguments to support it, so the belief is justified. Even if I never ask my neighbour to take care of my cat but it is reasonable to think that they would, I consider my neighbour trustworthy. Unforeseen circumstances can take place when my cat finally needs to be taken care of, such as my neighbour breaking their leg and not being able to access my house to take care of my cat. However, such an event would not make my neighbour any less trustworthy.

Based on the discussion above, I conclude that the key to trustworthiness is not what a trustee would do, but what would be reasonable to expect from them. From that claim I come to the following working definition:

**Definition 1.3.** *Trustworthiness* is the relational property of a trustee being worthy of a trustor's trust, in some particular domain. A trustee is trustworthy (for a trustor, in a domain) iff<sup>6</sup> it is reasonable for the trustor to delegate something (regarding that domain) to them.

I define trustworthiness as a relational property because in order to talk about a trustee's trustworthiness, there must be a trustor as well. A trustee cannot be worthy of trust if there is no trustor whose trust the trustee can be worthy of in the first place. According to the definition above, trustworthiness is a concept subordinated to the concept of trust. Thus, we see that trust and trustworthiness are intertwined.

---

## 1.5 The relationship between trust and trustworthiness

There is a normative connection between trust and trustworthiness. When trusting, the trustor assumes the risk of the trustee not behaving as expected, but they intend for such a risk to be as low as possible. The risk of the trustor being let down is lower when the trustee is trustworthy, so in principle, trustors should trust only trustworthy trustees. But the reality is different. It is common to encounter situations in which trust is misplaced. Trust is misplaced when the trustee is not worthy of it, but they are nevertheless trusted. A trustor can trust a trustee without having good reasons to do so, but this should not be the case precisely because it is an irrational practice. Thus, trust can exist without trustworthiness. If a trustor trusts an untrustworthy trustee, the trust of the former is unwarranted. This is called *mistrust*.<sup>7</sup>

<sup>6</sup>In definitions, I use *iff* as the abbreviation of 'if and only if'.

<sup>7</sup>For details about mistrust, distrust and other related concepts, see appendix [A](#).

If the trustor is overcautious, the converse can also happen: a trustor can distrust a trustworthy trustee if the trustor does not perceive the trustee as trustworthy. This happens when the trustor does not have enough information about the trustee, either due to their own ignorance or because the trustee portrays themselves in a misleading way. If my neighbour has never given me any hint that they are a cat person, I have no reasons to trust them with my cat. They are in fact good with cats, so the reasons to trust them exist, but such reasons are not available to me, so it makes sense to distrust my neighbour.

It is worth to be noted that both trust and trustworthiness are not binary concepts. They come in degrees. For example, imagine I consider other candidates to take care of my cat, besides my neighbour. It would be possible for me to trust all of them, which does not mean that I trust them to the same degree. I trust my mother more than I trust my neighbour, simply because she and I have a much stronger relationship based on a greater number of positive experiences. Trusting my mother more than my neighbour also means that, in my eyes, my mother is more trustworthy than my neighbour. In the end, I will end up delegating the care of my cat to one of them since, while trusting them to different degrees, I will to delegate the task to either of them. Unlike trust, the will to delegate is categorical: either I have the will to delegate or I do not. Different degrees of trust may all lead to delegation. But even though the end result is the same, trusting relationships can come in different shades.

So far, I have described the relationship between trust and trustworthiness. Next, I proceed to explore what distinguishes trust from reliance.

---

## 1.6 Trust versus Reliance

Now that I have intuitively characterised both trust and reliance, I proceed to carve out their differences. In later sections, this distinction will serve me to explore why it is more problematic to talk about ‘trust in AI’ than about ‘reliance on AI’.

In section [1.2](#), I defined reliance as the dependence of one party on another. I depend on a person or an object when, in order to achieve a goal, I need to delegate a task or role to them. The party that is being relied on can be either an object or a person. In the same sense that it is possible to depend on someone to achieve something, one can also depend on an object or a tool. Reliance is tied to action and delegation.<sup>8</sup> If there is no action to delegate, there is no reliance. This is a key point in which trust and reliance differ.

In section [1.3](#), I defined trust as a kind of relationship. I emphasised that trust is a relationship that can exist without delegation taking place. Unlike reliance, the key to trust is not delegation, but the possibility of it: the trustor believes that they could delegate something to the trustee if they ever needed to. Delegation may never become necessary, but the trusting relationship persists.

---

<sup>8</sup>As clarified at the end of section [1.2](#), when talking about relying on objects, I understand delegation in a broad metaphorical sense.

Most authors distinguish trust from reliance.<sup>9</sup> And many agree with Hawley (2014, p. 5) on conceiving trust as reliance plus some extra factor. The question is, what this extra factor is.

For those who propose motives-based theories on trust (Baier 1986; Jones 1996),<sup>10</sup> the extra factor would be the trustee's motivation to act. According to them, trusting implies taking into consideration not only how likely is the trustee to act as expected, but also what moves the trustee to act in such a way. On the contrary, for mere reliance the trustee's motives are irrelevant. For example, when I trust my neighbour to take care of my cat, I trust them not only because I expect them to do so, but also because I believe they have the right motives to take care of my cat. An adequate motive would be that they care about the well-being of my cat, or a general sense of responsibility. If I think my neighbour will take good care of my cat only motivated by the possibility of being paid for it, I may not trust them as much. On the other hand, reliance is based on expectation rather than motivation. Kant's neighbours relied on the behaviour they expected from him to adjust their clocks, but they did not trust him. Since Kant's motives to be by the Lindenalle every day at the same time had nothing to do with his neighbours, there was no trusting relationship between them; there was only merely reliance.

For authors holding non-motives-based theories of trust, the focus is on what the trustor believes they ought to expect from the trustee (Holton 1994; Hawley 2014). In this kind of frame, the difference between trust and reliance lies in the trustor's expectations, being higher when one trusts than when one merely relies on. Following this reasoning, trust differs from reliance in how the trustor perceives the trustee. In a trusting relationship, the trustee is seen as an agent who is able to hold responsibility; therefore, the trustor trusts them. When there is only mere reliance, the trustee is only a means to a goal. In consequence, the feeling of betrayal arises only when trust is broken, but not when reliance leads to failure.<sup>11</sup> Most accounts on trust fall into the category of either motives-based or non-motives-based theory. Either because of what motivates the trustee or what the trustor expects, both kinds of theories generally consider trust and reliance as distinct concepts.

Coming back to the conceptual difference between reliance and trust in AI, it is important to notice that people do not relate to other people in the same way as they relate to technology. Because of this, authors such as Nickel, Franssen, and Kroes (2010) or Ryan (2020) claim that it is not possible to trust AI, it is only possible to rely on it. As I said, talking about reliance on AI is relatively uncontroversial. However, in section 1.3 I highlighted features of trust that cannot be found in reliance. Nevertheless, some of those features are desirable in human-AI relationships. Trust is richer than mere reliance and, therefore, it is plausible that only humans can act as trustees. In the following chapter, I will explore the intuitive

---

<sup>9</sup>See Baier 1986, p. 234; Jones 1996, p. 4; Holton 1994, pp. 2–3; O'Neill 2002, p. 15; Hieronymi 2008, p. 215; McLeod 2021, p. 3

<sup>10</sup>Hawley (2014) distinguishes between motives-based and non-motive-based theories of trust. Motives-based theories take into consideration the trustee's motives to act as the trustor expects. In contrast, non-motives-based theories do not consider that the trustee's motivation plays a relevant role in trusting relationships.

<sup>11</sup>This is a shared point with motives-based theories: only trust can be betrayed, while reliance may be, if anything, disappointed (Baier 1986, p. 235).

notion of trust in AI and shed some light on why the general public and part of academia talk in terms of trust about human-AI relationships.

## 2

---

### *Kinds of trust*

---



---

## 2.1 Introduction

Artificial intelligence (AI) is making its way into our lives at a rapid pace, thus becoming a part of our social reality. As a novel technology, it offers many new resources and we use it to perform a variety of tasks, such as recognizing patterns, classifying images or generating speech in natural language. AI systems can process large volumes of data. From that data, they infer results which can be applied across various domains. In particular, AI can be used to make predictions or (recommendations for) decisions. These predictions may concern low-risk situations such as detecting spam emails or high-stakes situations, such as medical diagnosis. In the latter situations, the predictions or recommendations of AI can deeply affect people's lives. It is on this background that the question arises whether it is legitimate to *trust* the predictions or recommendations of AI systems, and whether it makes sense to employ the same concept as when we speak of people.

Because of the similarity of the situations (human-human interaction and human-AI interaction), it is common to talk about trust and distrust in AI. In particular, the notion of trust in AI has become key when analysing the role of AI in society and the consequences of its implementation. Indeed, trust appears as the central topic of numerous ethic guidelines, white papers and scientific publications.<sup>1</sup> When the results of a system are used in high-stakes situations, such a system should be worthy of trust, it seems. This academic discourse is based on the general use of the term 'trust'. In daily speech, it is common to encounter the intuitive analogy between human-human interaction and human-AI interaction. We conceptualize our relation to AI with the notion of trust, giving trust a normative character. It is generally thought that human-AI relationships should be based on trust since many of the situations in which AI is used are similar to situations where we ordinarily speak of trust in persons. Analogies are meant to shed light on issues that otherwise would be hard to grasp. Although the use of the notion of trust for the relationship between humans and AI might foster our understanding of this relationship, we should first ask whether this analogy is legitimate.

Humankind has been using technology to achieve different purposes for centuries. However, we do not tend to speak in terms of trust when we refer to technology. We do not say that we trust a tool (eg. a hammer) or a machine (eg. a car). If anything, technological artefacts are relied on. It is the user or the producer who we trust. As we have seen, trust is a richer concept than reliance. Intuitively, expecting someone to act according to our plans is enough to rely on them. Trust, on the other hand, seems to imply more than this. In ordinary language, the notions of trust and reliance are sometimes used interchangeably. By contrast, in the literature on trust, the two concepts are usually distinguished. It is widely agreed that reliance does not always amount to trust.<sup>2</sup> A person can rely on another one without trusting them and, in the same sense, it is possible to rely on objects we would not say we trust.

---

<sup>1</sup>Serve as examples OECD (2017), HLEG (2019), Leslie (2019), United Nations (2021), and Gunning et al. (2019).

<sup>2</sup>See Baier (1986), Jones (1996), Holton (1994), O'Neill (2002), Hieronymi (2008), and McLeod (2021).

When I get into a taxi and I trust that I will arrive at my destination, I trust the taxi driver, not the taxi. Given these remarks, we may ask whether trust is the right term to describe the relationship that humans ought to have with AI.

In this chapter, I discuss this central question. I will build up on the concepts of reliance, trust and trustworthiness to postulate the concept of ‘trust in AI’. By it, I refer to the kind of relationships that would be desirable to have between humans and AI. I will explore the possible parallelisms between this posit concept and more traditional accounts of trust in humans. The chapter is structured as follows. In section [2.2](#), I introduce a general notion of AI and the particular use of it that I will make during the rest of the thesis. In section [2.3](#), I focus on the aspects of AI that distinguish it from previous technology and that make AI a socio-technical tool. It is this aspect of AI that raises the question of whether it is possible to trust this new kind of technology. In [2.4](#), I propose case studies of typical scenarios in which AI is involved and, intuitively, should be the object of trust. From these cases, I will carve out the expectations for artificial trustees in these kinds of situations and compare them with what is expected from human trustees. On the basis of this comparison, I propose a tentative characterization of the posit concept of ‘trust in AI’. In section [2.5](#), I review the main literature on interpersonal trust. In section [2.6](#), I do the same for institutional trust. The goal of such reviews is to highlight the similarities and differences between trust placed in humans (either individual persons or institutions) and trust in AI. I conclude that the conceptual similarities between trust in humans and trust in AI weigh enough to label the latter as, indeed, a kind of trust. In section [2.7](#), I revisit the concept of ‘trust in AI’ as presented in section [2.4](#) and clarify some of the features that are expected in human or institutional trustees, but it is not so clear how they translate to artificial trustees. I will focus on the notions of motives, agency and (rational) reasons to believe that the trustee is trustworthy. The need to clarify how such concepts are to be understood when it comes to artificial trustees shows that the analogy between these different kinds of trust does not perfectly hold. Or at least, this is not the case if trust is characterized as previous accounts do. By pointing at such mismatches, I target the aspects of trust in AI that previous theories of trust are unable to reflect. Such aspects constitute the call for the development of a new account of trust that makes sense of this new kind of trustee.

---

## 2.2 AI as particular systems

Intelligence is often considered one of the defining traits of human beings. It can be defined in many ways. Broadly speaking, ‘intelligence’ refers to the capacity of processing information, assimilating it as knowledge and applying such knowledge to a certain context. Unlike other animals, humans are able of intelligent thinking, which translates into the ability to make abstractions, understand concepts and logically reason. But since the last century, it seems like such abilities may not be exclusively human anymore. We live in the age of artificial intelligence.

AI is the field devoted to developing technological models that aim to automate typical human activities, such as problem-solving, decision making or learning

(Bellman [1978]). Following Russell and Norvig ([2009], pp. 1–2), there are four main approaches to AI, depending what is considered to be the goal of AI: thinking humanly, acting humanly, thinking rationally or acting rationally.<sup>3</sup> In the last years, the scales seem to be tipping towards the latter approaches. Artificial intelligent systems are able to process large volumes of data that a person would never be able to. In this sense, AI surpasses human limitations such as time constraints and epistemic capacity, opening up a new world of possibilities.

AI is an umbrella term that encapsulates different sub-fields, such as natural language processing, computer vision or machine learning (ML). In this thesis, I focus on a particular kind of AI systems: deep neural networks (DNN or simply NN). A NN is a kind of ML algorithm that has gained notable popularity in the last years. NNs are able to process big amounts of data by passing the information through an intricate structure of nodes or neurons (see figure [2.1]). Nodes are singular computational units connected through a set of inputs and outputs that vaguely resemble human neurons (hence the name). The nodes are organized in layers used to screen the data. Data going through a layer of nodes means that some of the nodes will be activated. Activations in one layer determine the activations of the next layer, until they reach an output. NNs have a wide variety of applications and are already in use for assisting humans with various tasks. For example, OpenAI’s ChatGPT (Generative Pre-trained Transformer) is a language generation AI model that works using NNs. In this case, the model has been trained on large amounts of text data to generate human-like text. Users can feed text prompts to ChatGPT and the system provides seemly coherent answers to such prompts. To do so, each input (a text prompt) is decomposed into its parts, which in this case are individual words. Then, the NN weights the ‘importance’ of each word as it goes through each layer of the system. This way, ChatGPT is able to generate a context-aware representation of the input prompt. Then, this representation passes through the fully connected neural network layer, where the model makes a prediction that comes out as an output. In other words, ChatGPT syntactically analyses the users’ sentences and calculates which sequence of words is most likely to be an appropriate response to the input sentence. This way, the system simulates having a conversation with the user in natural language (see figure [2.2]).

For the purposes of this thesis, when I talk about AI, I am referring to ML algorithms and, in particular, to NNs. Given how broad the AI field is, I have chosen to narrow it down by focusing on this kind of system due to pragmatic reasons.<sup>4</sup>

---

<sup>3</sup>Russell and Norvig themselves fall in the category of authors who consider AI as the discipline devoted to developing systems that act rationally. There is a rich debate among those who defend that AI systems aim to *think* rationally (Russell and Norvig [2009]) and those who consider that the aim is to *act* rationally (Bellman [1978]). For the purposes of this thesis, it is not necessary to adopt a strong position regarding the conceptual goal of AI. Therefore, I won’t engage further in this debate.

<sup>4</sup>Besides, the discussion could be easily extended to other methods. An alternative example to NNs is logistic regression models. Such models are easier to interpret than NN since they work with linearly separable data. However, they still pose a problem for trust. In ML, logistic regression is used to predict the likelihood of an outcome, such as how likely a tumour is to be malignant, for example. In high-stakes situations of this kind, predictions should be able to be trusted. However, since the predictions are made by non-human entities, the question of whether is legitimate to use the term ‘trust’ holds.

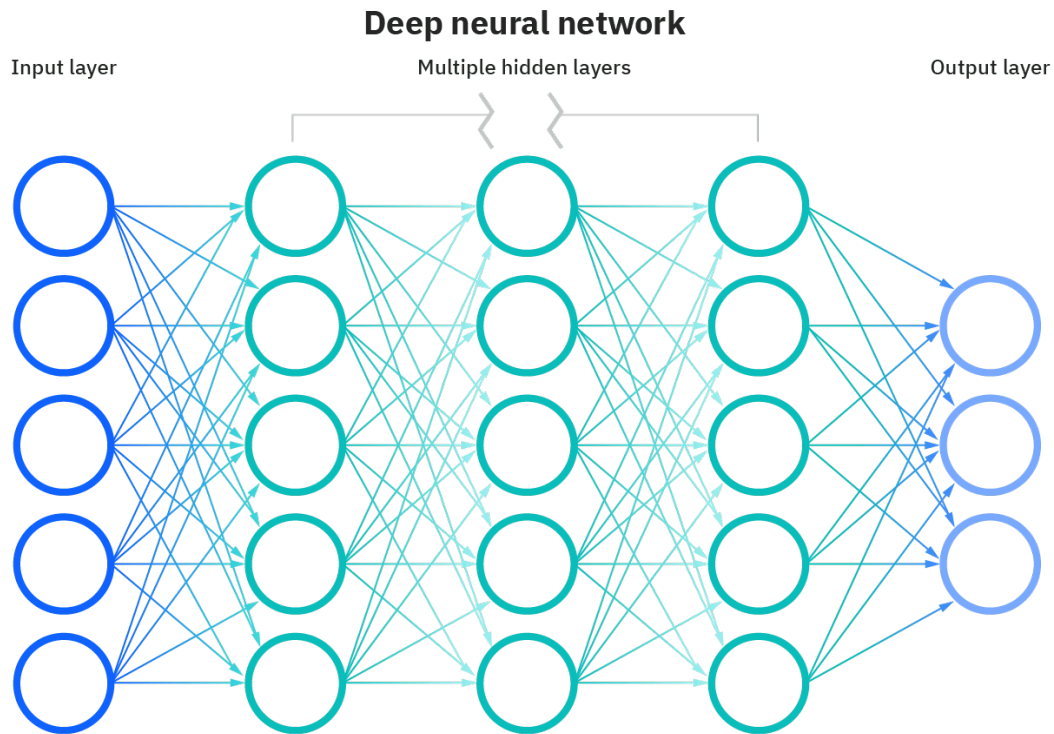


Figure 2.1: NNs consist of a large number of nodes or neurons organized in layers. Data goes through this structure, activating different nodes in each layer. Activations in one layer determine the activations of the next layer, until they reach an output (image source: IBM Cloud Education (2020)).



Hey, how do you work?




I am a language model developed by OpenAI. I use machine learning algorithms to generate text based on patterns in large amounts of text data. When I receive a prompt, I analyze it and generate a response based on the patterns in the data I was trained on. I was trained on a diverse range of texts, so I can respond to a wide variety of questions and generate text on a wide range of topics. 

Figure 2.2: Example of text response generated by GPT-3 using OpenAI's chat (<https://chat.openai.com/chat>).

NNs have gained notable popularity due to their high rate of accuracy. By accuracy, I mean that NNs mostly arrive at satisfactory results in their used contexts. For example, sticking to the case of ChatGPT, the system is considered technically successful if it is able to provide seemingly coherent answers to most<sup>5</sup> of the inputs it receives.

NNs offer a wide range of possibilities, both in low and high-risk situations. When I talk about high-risk situations, I refer to scenarios in which the outcome provided by an AI has an important impact on people's lives. This is the case for medical diagnostic or loan management systems. Unfortunately, given their high complexity, NNs are what is commonly referred to as 'black boxes' algorithms. Black boxes are opaque AI systems, meaning that it is not possible for human beings to understand how the system reaches its outputs. The black box problem poses an interesting philosophical issue. The trade-off between accuracy and transparency raises the question of whether a system whose workings are not fully understood ought to be trusted, even if it offers high rates of accuracy. It is this trait that makes NNs a good candidate to explore the questions this thesis addresses, namely: is trust a suitable term to describe human-AI relationships, and if so, what does trusting AI mean?

---

## 2.3 AI as a socio-technical tool

### 2.3.1 Users, tools and systems

Tools, broadly understood, are means that aid in accomplishing a task. In this sense, the term 'tool' can include a wide variety of devices and machines, such as a ladder that is used to reach a lamp or a car that is used to arrive at a destination. AI is a tool too, whilst it can be used to fulfil several purposes, like data classification or prediction making. However, AI is not a tool like any other.

Generally, tools are used in a passive way. The user of the tool is considered the performer of whatever task the tool is being used for, while the tool is only a means. In the case of AI, the tool adopts more of an active role. Artificial intelligent systems *learn* from data and arrive at results that could not be predicted by looking only at the system's design. Thus, AIs contribute to the goals that they are used for to a much greater extent than any other kind of tool. AIs *produce* their own outcomes. Humans engage with those outcomes, rather than merely *using* the systems to reach outcomes by themselves. This is what makes trust a **socio-technical tool** (for short, STT).<sup>6</sup>

---

<sup>5</sup>Here, I use 'most' in a loose way. There is no exact percentage of successful cases from which a system is considered technically successful. What is considered a high rate of success varies depending on the system to rated and its application.

<sup>6</sup>In this thesis, I use STT as an abbreviation of 'socio-technical tool'. Not to be confused with 'Science and Technology Studies', 'Science, Technology and Society' or 'Social Studies of Science and Technology' (Van House 2003). I use the term 'socio-technical tool' instead of 'socio-technical system' because, to my understanding, a socio-technical system consists of wider processes than the ones I aim to point at (see the characterization of socio-technical systems offered by Van House

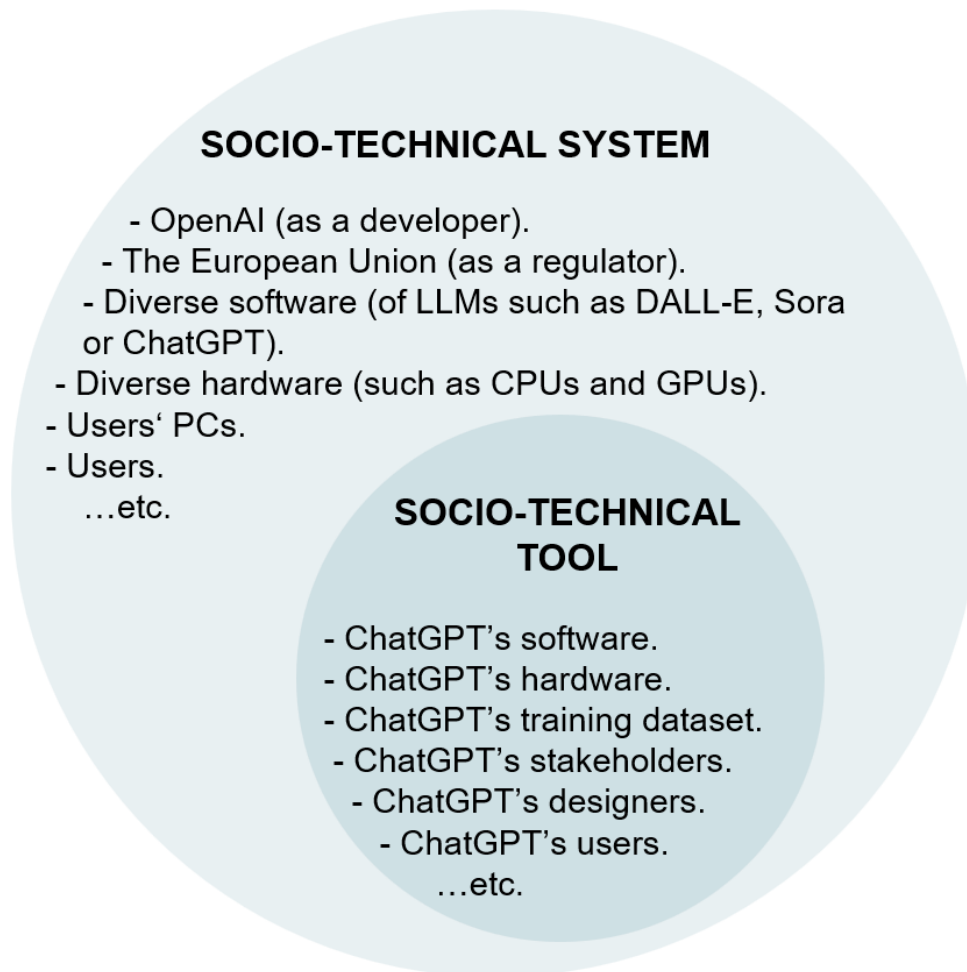


Figure 2.3: ChatGPT as an example of a socio-technical tool integrated within a broader socio-technical system.

A STT is a system which consists of human and non-human agents who interact for various purposes.<sup>7</sup> I consider AI a STT because, as stated above, humans *interact* with AI rather than merely use it. AI involves several components, and not all of them are technical. An AI, like many other artefacts, is composed of hardware and software. Both of them constitute the technical part of the system. In order to function, AI also needs data and human interaction. This is the social part. A system's performance depends not only on its material pieces and code but also on the data that it is fed with and the trial and error it goes through. In this sense, artificial intelligent systems suppose a network of technical and social elements that contribute to producing results. Such results are not fixed achievements. On the

(2003, p. 72)). For example, a book publishing system would be a socio-technical system that includes authors, publishing companies, printers, data-bases, etc. By socio-technical tool, I refer to specific models and devices, such as OpenAI's ChatGPT, for example.

<sup>7</sup>This definition is inspired by Simon's (2010) concept of socio-technical epistemic system.

contrary, they can improve over time, depending on how the AI continues to be used.

This way of relating with the systems conceptually makes them something more than technical artefacts. The role of AI in society goes beyond the role that previous technology played. To elaborate on this idea, the Actor-Network Theory (ANT) comes in useful.<sup>8</sup> ANT is a method in science and technology studies that abolishes the essentialist distinction between society and technology. According to this idea, society would be an “ongoing achievement”, rather than an entity that can influence and be influenced (Callon (2004), as cited in Simon (2010, pp. 72–79)). When society is understood in this dynamic way, it can be considered that AI actively participates in it. It does so by contributing to several social aspects, such as the access and processing of information, mobility or communication.

Nevertheless, AI is not the first technology with a clear social effect; serve as examples the television, the telephone or the car. But following up on footnote <sup>6</sup>, I consider artefacts such as a car or a telephone technical tools, even though they can be part of socio-technical systems (therefore their social aspect). In contrast, AI is not only part of a bigger socio-technical system: it is a socio-technical tool on its own (see figure 2.3). The difference is subtle, but relevant. Artefacts such as the telephone have a social component whilst they are part of a bigger system. Such a (socio-technical) system includes not only the artefacts, as the technical part, but social practices, which in the case of the telephone would be communication. The social component of AI goes a step further. Besides the bigger system to which it belongs, an individual AI system already possesses a social component since it relies on data and human interaction to function. Humans use AI, but in a way, AI uses humans too. While a telephone can exist with no one using it, without human data and interaction, AI is not conceivable.

### 2.3.2 Quasi-agents

The novelty of AI creates a conceptual gap between this kind of system and other technological artefacts. The situations and applications of AI differ significantly from previous technologies, as AI possesses an unprecedented ability to replicate human decision-making processes more closely than ever before. Furthermore, the outcomes or predictions generated by AI systems can be unforeseeable even by their designers. This has led some authors to discuss these systems as having (*quasi*-)agency (Coeckelbergh (2009, p. 184). According to Coeckelbergh (2009), humans treat non-humans *as if* they were agents in order “to support the (quasi)moral and social dimension of our dealings with them” (p. 184). By ‘quasi-agency’, I refer to this kind of ‘as if’ approach. Such an approach allows us to make sense of AI’s workings by conceptualising them as ‘decision making’, for example. This and similar terminology is used as an analogy to human activities. It is known that, at the moment, AIs lack consciousness, freedom and other traits which are necessary to make decisions. Nev-

---

<sup>8</sup>For further details about ANT, see Callon (2004).

ertheless, both the public and academic literature talk about AI’s decisions, referring to artificial intelligent systems *as if* they were agents.<sup>9</sup>

In this section, I have conceptualised AI as a socio-technical tool. In this sense, an AI system is conformed not only by hardware and software, but also by human input and data, from which the system ‘learns’. In addition, what I mean by socio-technical tool is that, unlike mere tools, the use of AI systems goes beyond extending a particular user’s ability. AI systems –particularly machine learning (ML) ones– *seemingly* ‘master’ abilities themselves, conducting apparent decision-making. By this, I do not mean that AI systems have (full) agency. In any case, some sort of *quasi-agency* (Coeckelbergh 2019) could be attributed, more as a result of how users engage with the systems than of the systems’ abilities per se. Hence, such an appearance of agency (or *quasi-agency*) gives rise to the intuition of anthropomorphising some AI systems and engaging with them in a similar way in which we interact with other humans (Placani 2024). One reason for this intuitive anthropomorphisation is that the user does not perceive the mechanisms through which the system mimics human skills, but only the appearance of such replicated skills. Even if the user is familiar with how AI works, they still interact with an artefact that somehow mimics human behaviours. Therefore, the natural response is to react to such behaviours as one normally does with fellow humans. Soon, I will argue that this dynamic of mimicking and engaging with mimicked behaviours includes trusting relationships.

AI systems exhibit an apparent ability to achieve results owing to its capacity to identify patterns and derive solutions (relatively) autonomously, beyond the explicit programming by its developers. I qualify this ability as *apparent* because, as mentioned, humans interact with AI systems *as if* they were fellow humans when such systems succeed at mimicking typical human skills. Therefore, whether AI systems achieve results autonomously or not is beside the point: as long as they are perceived as autonomous, they will be perceived as potential trustees.

As specified in the previous section, in this thesis, I focus on NNs. The goal of NNs is to minimize their loss function; that is, to make predictions as accurate as possible, balanced against the risk of overfitting, particularly over the training data set. In adjusting its weights, the neural network follows a predetermined algorithmic process, primarily through backpropagation and gradient descent. While this process is deterministic, meaning it follows a set of defined rules and calculations, it is important to note that the designers do not explicitly set the final weights. Instead, the network iteratively adjusts its own weights based on the data it processes. This gives the impression of autonomy, but it is within the confines of the algorithm’s structure and the data it is trained on. Therefore, the ‘freedom’ of the system lies in its ability to autonomously find the best solution within the parameters set by its algorithm, rather than the designers specifying weight values directly. Thus, how we use and relate to AI differs from how we relate to previous technology.

Unlike previous technology, AI uses data to achieve results that are unpredictable even for its designers. This is so because, even if AI systems are not ‘free’ to pursue autonomous thinking as a person would, the system is able to iteratively

---

<sup>9</sup>Deeper discussion on AI’s agency goes beyond the scope of this thesis. For those interested, an in-depth discussion of forms of non-human agency can be found in Barad (1998), Suchman (2006), and Ågerfalk (2020), for example.

adjust its weights based on the specific dataset and the guiding principles of back-propagation. Because of this, even if AI ‘learns’ in a significantly different way from humans,<sup>10</sup> its way of managing data poses a novelty that requires its users to engage with the technology in a novel way. Unlike previous technology, AI systems do not give us the means to merely extend our own abilities: it *seems* to exhibit abilities of its own. Therefore, the user does not merely use what the system provides, but they interact with the results, continuing to feed the system with new data and making its results better over time.

Its state of quasi-agents makes AI’s interaction with humans novel and its social role greater. Thus, the fact that AI is not a mere tool, raises the question of whether it can be the object of more than mere reliance. Here is where the concept of ‘trust in AI’ comes into play.

---

## 2.4 Trust in AI

In this section, I argue that trust is a suitable concept to describe relationships between humans and artificial intelligent systems. My argument runs as follows: In [2.4.1](#), I will discuss typical scenarios<sup>11</sup> involving AI that require trust in a normative sense. In [2.4.2](#), I aim to describe the posit concept of ‘trust in AI’. To do so, I will carve out the common features from the scenarios discussed and break down the normative expectations that arise in them. In [2.4.3](#), I will show that what is expected from AI in such cases is analogous to what is normally expected from a human trustee.

### 2.4.1 Typical scenarios involving artificial trustees

#### Case A: An example from the medical domain

Alice presents a gradual lack of audition and loss of her equilibrium. Her doctor suspects that she may have a benign brain tumour known as vestibular schwannoma (VS). The main treatment for this tumour consists in extirpating it through surgery. In order to confirm or discard this diagnosis, Alice gets an MRI done. The MRI goes through a brain tumour detection and classification neural network (NN). In order to incorporate explainability in the early stages of the process, Grad-CAM is implemented on it. Grad-CAM is a popular technique used to highlight the areas where an NN model pays attention in order to classify images; that is, which pixels are relevant for the system in order to arrive at conclusions. The NN confirms what Alice’s doctor suspected: she has VS. Examining the Grad-CAM’s analysis allows us to see why the system concludes so since it highlights the area of the MRI where the tumour can be observed.

---

<sup>10</sup>It is not yet fully known how human minds work and how we learn entirely. However, what it is known, is that we do not use methods such as backpropagation.

<sup>11</sup>Case A and case B are toy examples inspired by [Windisch2020ImplementationSlices](#) and Ortega et al. [\(2021\)](#) respectively

### Case B: An example from the recruiting sector

Bob and Betty apply for the same job vacancy as accountants for a company called Baum. Their curriculums are identical: same education, same work experience, same age, etc. The only varying feature is their gender. The hiring process is held by an automated recruiting system, in which PRIDE has been implemented for explanatory purposes. In this case, the recruiting system chooses the profile which fits the vacancy best, and PRIDE is able to tell which weight each feature played in such a decision.

In the end, Bob is hired for the job and Betty is rejected. When Betty asks for feedback about her application and what she could have done differently to get the job, Baum checks PRIDE's analysis. PRIDE shows that if *gender*(0) appears more often than *gender*(1) in the recruiting system's scoring rules, then *gender*(0) would score higher. Then, the answer is that there is nothing Betty could have done differently, since she cannot change the fact that she is a woman. Nevertheless, this was the feature that got her disqualified from the job against Bob.

### 2.4.2 A tentative description of 'trust in AI'

The examples from the previous section expose scenarios in which trust in the corresponding system is required. Both diagnosis and recruitment processes are high-stakes cases that deeply affect people's lives. Therefore, it would not be responsible, maybe not even feasible, to use systems which are not trusted in scenarios of that kind. In light of the case studies above, I provide tentative features and expectations to describe the posit concept of 'trust in AI':

#### (1) A relationship with positive expectations

'Trust in AI' is a kind of relationship between a person (the human trustor) and an artificial intelligent system (the artificial trustee) in which the human trustor has positive expectations about the system's performance. In cases A and B, such positive expectations would be a successful diagnosis and the ideal candidate for the job position, respectively.

#### (2) Need or wish for delegation

Human trustors aim to establish trusting relationships with artificial trustees when they need or wish to delegate certain kind of tasks to those trustees. In high-stakes situations, like cases A and B, trust is required in order to use the system. It would be irresponsible to use AI that affects people's lives in such a significant way if the system cannot be trusted.

#### (3) Risk of failure

When a human trustor 'trusts' an artificial trustee, the trustor makes themselves vulnerable by willingly accepting the risk of failure in the system. Trusting a system implies expecting that such a system will successfully accomplish the task it was designed for. However, human trustors are aware that systems are not infallible. For example, in case B, PRIDE reveals that the candidate's

gender was a parameter arbitrarily taken into account to select the best person for the job. Since that is not an appropriate criterion in a recruiting selection process, the system may not successfully fulfil the task, and even if it did, its performance would be ethically questionable because it used a biased procedure. When trusting AI, human trustors are aware that these kinds of issues may happen, even though the aim is to avoid them.

#### (4) **Motives-based**

As pointed out in (2), trusting relationships are built on the possibility of delegation, which will materialize in case the trustor wishes or needs to do so. Thus, trust goes *beyond* successful delegation, which makes it distinct from reliance. As I showed in section 1.6 from chapter 1, trust is a richer concept than reliance. Reliance is more closely tied to the act of delegation, whereas trust primarily concerns the potential for delegation. In the case of AI, a human trustor successfully delegates a task to an AI system if the system produces an accurate result for that task, using methods approved by the trustor. Thus, the process that the system follows to achieve results is relevant, as case B exemplifies. Imagine that, after all, Bob happens to be the best candidate for the job. If that were the case, then the system would have found the best candidate for the job, that is, it would have provided an accurate prediction. However, even in this scenario, after the gender-bias is known, the system should not be trusted. Given the socio-technical status of AI, not only technical aspects should be taken into account for its successful implementation in society. It is not only relevant what outputs can an AI system provide, but also why it provided such outputs. In this sense, trust in AI should be motives-based: the *motives* that lead a system to perform as it does play a central role when determining whether the system oughts human trust.

#### (5) **Rational reasons**

‘Trust in AI’ must be grounded in rational reasons. That means that it should be reasonable to believe that:

- The human trustor’s motivation is aligned with the artificial trustee’s (the trustor’s motivation is reflected in the system’s process to arrive at results).
- The artificial trustee has the competence to act accordingly to such motivation.
- Therefore, the artificial trustee will most likely succeed at the task that the trustor may delegate to it, if the trustor wished or needed to do so.

By rational reasons, I mean those that can be legitimately used to justify trust and, therefore, have the potential to be agreed on by others than the trustor. In case A, a rational reason to believe that the NN would successfully perform would be the high success ratio of the system. Another rational reason to believe

that the system's motives align with the trustor's could be the availability of explanations that allow human trustors to understand how the system works and why its workings lead to successful results. This was not the case in case B, and therefore the system should not be trusted.

### (6) **Intersubjectivity**

As a consequence of (5), 'trust in AI' should be intersubjective. In chapter 1 (section 1.3), I said that trust is subjective because it is usually based on personal judgements and experiences. However, in the case of AI, trust should overcome such subjectivity. That is, if a system is considered to be trustworthy, it should be a general consideration rather than a personal judgement by a particular trustor. This normative demand comes from the fact that successful implementation of AI involves not only technical aspects but also social ones. It would be beneficial to discuss in a social frame the reasons to trust or distrust AI. By taking a variety of angles into account, the acceptance of AI in society would be better contrasted, more democratic and a further scrutinized social position. Following case A, an AI diagnostic system could not be implemented in a hospital if only a small group of practitioners trusted it, and their trust was based on personal experiences that cannot be generalized. In the field of AI, trust demands a certain grade of objectivity; or at least the possibility of rational evaluation of the reasons to trust. This possibility of rational evaluation is what I designate as *intersubjectivity*.

It is worth noting that points (1)-(3) are descriptive, while (4)-(6) constitute normative expectations. That is, (1)-(3) describe how some human-AI relationships are: what human trustors expect, when those relationships are established and the risk that trustors are aware of. Incidentally, in (4)-(6), I described how 'trust in AI' *should* be.

For the most part, chapter 1 was descriptive. In it, I gave a general characterization of trust: what the trustor expects and what they would be willing to do. When exploring the relationship between trust and trustworthiness in section 1.5, I hinted that trust is only warranted when placed on trustworthy trustees. Normatively, trust should be warranted, so trustors should place trust in trustees that are worthy of it. Let us see now whether trust in humans can be described in a similar vein in which I have tentatively characterized 'trust in AI'.

### **2.4.3 Trust in AI and trust in humans: shared features**

Let me begin this section with a third case study. This time, the example involves only human parties.

#### **Case C: An example from daily life**

Claire is going on holiday and she needs someone to take care of her cat. Her neighbour Christian has a cat himself, which he has talked to Claire about several times. Christian is available during the dates Claire is gone. Christian has taken care of Claire's cat before, so once again, Claire asks Christian the favour, which he accepts.

Cases A and B led to the formulation of the remarks and normative expectations listed in 2.4.2. Now, I will argue that such remarks are not exclusive to artificial trustees (or at least they should not be). On the contrary, those features of trust can be also found in cases in which the trustee is a person. Making use of the case studies above, I proceed to illustrate such a claim.

Points (1)-(3) describe ‘trust in AI’ in a way that does not conflict with the general definition of trust that I provided in section 1.3. Generally speaking, trust is a kind of relationship in which the trustor believes that they could delegate something to the trustee (because of the positive expectations the former has about the latter). The trustor is aware of the risk that the trustee may not act or perform as the trustor wants. However, the trustor considers such a risk minimal because of their positive expectations. Therefore, if the scenario in which delegation becomes necessary (or desired) materialized, then the trustor would indeed delegate that something to the trustee. This describes situations in which the trustee could be either a person or an AI.

In point (4), I highlighted the fact that ‘trust in AI’ should be motives-based. That means that trust in this kind of systems should go beyond accuracy, which distinguishes the concept from mere reliance. This is necessary for AI to comply with ethical standards, but it is a normative demand that applies to human trustees too. In cases of trust in people, the equivalent of accuracy would be successful delegation. In case C, a successful delegation would mean Claire’s cat being properly taken care of. Imagine that Claire, instead of following a logical reasoning, trusted Christian just because of how he looked. Nevertheless, luck is on Claire’s side and Christian happens to take good care of her cat. Even if Christian acted how Claire expected, Claire’s trust is misplaced.<sup>12</sup> Trust is not only about relying on someone to do something. Trust involves the motives or reasons for that person to act, and the whole network of beliefs that makes the trustor trust. This also applies to cases in which the trustee is an AI.

Point (5) alluded to the fact that trust should be grounded in rational reasons. This is a must when trusting AI, and it is clear for both cases A and B: there must be rational reasons that support the use of AI in domains such as medical diagnosis or recruiting. Otherwise, we accept a risk that is too high without justification. That is undesirable, both for ethical and pragmatic reasons. When the stakes are high, the trustor should be able to justify their trust in case things go wrong. This is the case for situations that involve artificial trustees, but also for human ones. Back to case C, I have spelt out some of the (rational) reasons that Claire has to trust: Christian is good with cats and he has taken good care of Claire’s cat before. This makes Claire’s trust rational. Alternatively, Claire could irrationally trust Christian just because he happens to have a sweet face, and she has the feeling that he would never harm her cat. That irrational form of trust is one form of human trust, but it is not the only one (and in any case, it is not how trust *should* be). The higher the stakes are, the more important it is to have strong arguments to support trust. This is key in institutions and enables social cohesion. For example, we trust doctors or judges. This trust is far from being irrational: there is a whole system that endorses

---

<sup>12</sup>See appendix A

the competencies of those figures. Both when it regards human or artificial trustees, having rational reasons to trust is desirable since it will be more likely that trust is well placed.

Finally, in point (6), I stated that ‘trust in AI’ should be intersubjective, meaning that, since trust should be grounded in rational reasons, it should be possible to agree on those reasons. When it comes to trusting people, that should be the case too (even though, people do not always place their trust following logical reasoning). In case C, Claire trusts Christian. The personal history between the two probably plays a role in their trusting relationship, but Claire has further reasons. If their neighbour Charlotte had exactly the same information that Claire has, it is likely that she would also trust Christian. Christian has a cat which he takes good care of. Then, the trust placed in Christian is not purely subjective but intersubjective, since both of his neighbours share rational reasons to agree into trusting him. Overcoming subjectivity is desirable both when trusting humans and when trusting artificial intelligent systems.

From this section remarks, I conclude that what is expected from artificial trustees is not so different from what is expected from human trustees. By this, I do not mean that trust in humans finds a perfect homologous in ‘trust in AI’. Trusting people is different from trusting AI in many ways, such as the human trustees having intentions or the possibility of betrayal, for example. However, interpersonal trust is also different from other kinds of trust, such as institutional trust, self-trust or trust in science. Trust is a rich concept that goes beyond the most paradigmatic cases of interpersonal trust. Because of this, I will argue that trust is a suitable concept for describing the relationship that humans should have with AI.

In the next sections, I will review the main literature on trust. I have divided the corpus of theories into two groups: accounts of trust that concern trust placed in persons (section 2.5) and accounts that focus on trust placed in institutions (section 2.6). Usually, these kinds of trust are referred to as interpersonal and institutional, respectively. My aim is to analyse whether posit features (1)-(6) can be found in the existing literature on interpersonal trust. If the answer is positive, it will strength the argument of ‘trust in AI’ being indeed a kind of trust. In case the answer is negative, it will constitute an argument against the use of the term ‘trust’ to describe human-AI relationships.

---

## 2.5 Interpersonal Trust

Interpersonal trust refers to cases of trust involving single persons. It is widely agreed that interpersonal trust is the dominant paradigm of trust (McLeod 2021). There is a vast body of literature on the subject of interpersonal trust,<sup>13</sup> but no clear consensus on how to characterise it. I have divided the main accounts into three categories:

<sup>13</sup>See Baier (1986), Baker (1987), Jones (1996), Hardin (2002), Hawley (2014), Simon (2020), and Nguyen (2019).

accounts that portray trust as risk-assessment, accounts that describe trust as an attitude and accounts that characterize trust as a contract.

### 2.5.1 Trust as risk-assessment

“Risk-assessment views” is a term taken from Jones (1998). It refers to positions that analyse trust as a way to avoid risk. When trusting a person, there is always the risk of that person letting us down by not behaving as we anticipated. According to risk-assessment theories, if the trustor thinks that the trustee will most likely behave as the trustor expects (the perceived risk of being let down is sufficiently low), then the trustor trusts. Risk-assessment views are motives-based theories, even though the specific motives that move the trustee are not important. What is important is whether the trustee’s motives will (most likely) make them act as the trustor expects. Examples of this kind of account can be found in Gambetta (1988) or Hardin (1993).

According to Gambetta (1988), we trust merely to obtain what we want. Under this view, trusting is a way of achieving our goals assuming the least risk possible, taking into account that our trustees are as free to betray us as we are to betray them. Gambetta even suggests the possibility to parameterize trust, from 0 to 1, where 0 corresponds to distrust and 1 to complete trust. Then, 0.5 would correspond to the point at which the trustor does not know at all what the trustee will do. According to this theory, what motivates the actions of the trustee is not important. The only thing that matters is whether such motivation results in fulfilling the expectations of the trustor, and how likely the motivation is to endure.

Hardin (2002) proposes a theory of trust based on self-interest, known as the “encapsulated interests” account. According to this theory, a trustee encapsulates the trustor’s interests when the trustor’s interests are included in the trustee’s motivation to act. In Hardin’s (1993) words, “I trust you because I think it is in your interest to take my interests in the relevant matter seriously [...] That is, you encapsulate my interests in your own interests” (p. 1). Following this reasoning, trusting trustees who encapsulate the trustor’s interests minimizes the risk for the trustor of being betrayed. Thus, the key to trust is that the interest of the trustor and the trustee are in alignment. That is, both the trustor and the trustee benefit from the trustee’s actions in some significant way. As in Gambetta, it is irrelevant what exactly these interests or motives are, just that they are shared between both parties.<sup>14</sup>

Both Gambetta and Hardin propose accounts in which trust can be quantitatively evaluated. For Gambetta, this quantitative evaluation lies on how certain the trustor is about the trustee’s future behaviour. For Hardin, the criterion would be the degree of alignment of the trustor’s and the trustee’s interests. The strength of risk-assessment views is that they portray trust as reasonable. I understand that trust is reasonable if it is based on logical reasoning; that is, if there are rational

---

<sup>14</sup>Note that Hardin’s view of trust ignores cases in which trustor and trustee do not aim towards the same goal, and nevertheless they trust each other. An example of such adversarial situations (Bordt et al. 2022) would be job recruitment, like case B from section 2.4.1 illustrates. The recruiter aims to find the best candidate for the job, while the applicant aims to get the job, regardless of whether they are the best candidate or not. Still, the recruiter trusts that each applicant is honest in their interview and would do their best to fit the position, and the applicant trusts that the recruiter carries a fair selection process.

reasons to trust. By rational reasons, I mean those that can be legitimately used to justify trust and, therefore, have the potential to be agreed on by others than the trustor. According to risk-assessment views, the trustor trusts because

- a) The trustor considers it is probable that the trustee will act as the trustor wants (Gambetta), or
- b) The trustor thinks that it is in the trustee's interest to act as the trustor wants (Hardin).

Both reasons above would qualify as rational. Using probability to predict the future and placing trust according to that probable future is logical reasoning. Similarly, it is logical to think that if someone shares my interests, they will behave as I want them to. If it was the case that trustors trusted attending only to rational reasons, intersubjective discussion on who is worthy of trust would be much simpler. This characterization of trust matches the demand for rational reasons that human trustors typically expect from artificial trustees (as pointed in section 2.4.2).

The main disadvantage of risk-assessment accounts is that they fail to distinguish between trust and reliance.<sup>15</sup> The reason is that, even though they are motives-based theories, the specific motives of the trustee to act are not addressed. Reliance concerns strictly delegation. I rely on someone to do something when I simply want that something to be done. Why the person did it does not really matter. This is precisely how risk-assessment theories define trust. According to these theories, the trustor's criterion to trust the trustee is that the risk of the trustee acting against the trustor's desires is low enough. The reason for that risk to be low is unimportant. If what motivates the trustee to act is ignored, then trust equals reliance. However, as pointed in section 1.6, the difference between trust and reliance is meaningful. I may *rely* on Kant to adjust my watch, no matter why he happens to be at the same place every day at the same time. However, if I *trusted* Kant for bringing me medication on a daily basis, I would care that his motive for being punctual is how much he values my well-being, and not just that he happened to be there. Both situations are not equivalent, and neither are the consequences of Kant failing to be on time. As we have seen, the difference between trust and reliance is relevant both when the trustee is either a person or an AI system. On this basis, I conclude that risk-assessment views of trust are incomplete.

There is a further critique to these accounts. If we accept risk-assessment accounts on trust, the trustors should trust in situations in which, in reality, not even reliance would take place. According to risk-assessment views, trustors trust as long that the risk of being let down is low enough. Thus, depending on the particular theory, the criterion to trust lies on whether the trustor and the trustee share the same goals (Hardin) or on how likely it is that those goals are reached (Gambetta). Let me illustrate how this fails using case C once again. According to risk-assessment views, Claire will trust Christian to take care of her cat as long as the risk of her cat being unattended is minimal. This way, Claire should trust Christian regardless of

---

<sup>15</sup>Like other authors (Jones 1998; Hawley 2014; Goldberg 2020), I consider trust to be different from reliance. In particular, Jones (1998) criticises the lack of distinction between trust and reliance in risk-assessment approaches to trust.

his motives to take care of Claire's cat. So it should not matter whether Christian's motivation has to do with his love for cats or whether he has the constant hallucination that the cat is actually his deceased daughter. But actually, no one in their right mind would trust a neighbour who cannot distinguish reality from their own hallucinations. No matter if the well-being of Claire's cat interests them both due to such hallucinations, or if the hallucinations are the reason why it is probable that the cat is treated like a king. Then, what motivates the trustee's actions seems to matter for the trustor.

I consider trust to be a rich concept and the relationship between trustor and trustee more than a means to end connection. In a trusting relationship, the trustee acting according to the trustor's desires is important. However, it is not the only fact on which trusting relationships are sedimented. Following the discussion above, it is also important that the trustor deems appropriate the trustee's motivation to act. In fact, two people can hold a trusting relationship even if there is nothing to delegate, just the hypothetical possibility of it. This way, the trustor will trust the trustee if the trustor believes that, in a hypothetical scenario, the trustee would act in a certain way moved by certain motives. Such a scenario may never materialise. But even if the hypothetical scenario never becomes actual, the trustee's motives hold and with them the trusting relationship.

Risk-assessment theories do not capture the aspects above, and present trust as some sort of stag hunt game. I proceed to develop the resemblance. In *Discourse on the Origin of Inequality*, Rousseau (1755) presents the case of two hunters. Let me call the hunters Hans and Hannah for clarifying purposes. Each of them must individually decide whether to hunt a stag or a hare. They know that two people are needed to hunt a stag. If each of them hunts a hare, cooperation is not needed and both hunters keep their autonomy. However, a stag is much more valuable than a hare. Then, Rousseau concludes, the best option for each hunter is to give up total autonomy in the interest of both of them. Connecting this story with the topic that occupies us, Hanna trusts Hans based on the risk of him acting against her benefit. Then, Hannah only trusts Hans because she knows that hunting the stag also benefits Hans, and she expects him to act according to his interests. Or, Hannah may reason, the most likely scenario is that Hans holds the same hypothesis about her and therefore he will trust her back. Their interests are aligned. Then, the risk of being betrayed by him is low enough to trust. I consider that trust goes beyond this kind of relationship since, as I said, it can take place even if there is no stag to hunt (but a hypothetical one).

In addition, seeing trust as risk-assessment implies considering it a conscious decision. This feature seems to be desirable when the trustee is an AI, but in interpersonal trust that is not always the case. We often wish we could trust someone, but we cannot. We cannot simply decide how we feel about someone, which image we have of them or what kind of relationship we establish with them. Relationships are built over time and experience. The subjective probability of a trustee behaving a certain way is part of what makes a trustor (dis)trust, but trusting relationships are more than that. The personality of the trustor also plays a role: what they value in a trustee, how gullible they are and so on. With the aim to capture nuances of trust that go beyond mere self-interest, other authors define trust as an attitude.

## 2.5.2 Trust as an attitude

There are many reasons why we trust someone, and not all of them can be reduced to self-interest. An attempt to capture this consists in focusing on the qualitative side of trust. That is, understanding trust as an attitude we have towards someone, rather than as a quantitative evaluation. I have chosen Baier (1986), Jones (1996) and Nguyen (2019) to illustrate the wide spectrum of approaches to attitudinal trust.<sup>16</sup> This choice is motivated by the fact that, while the three papers refer to trust as an attitude, they are significantly apart in time. That allows me to contrast long-established views with a more contemporary one.

Baier (1986) characterizes trust as an *attitude of accepted vulnerability*. In her own words, trust is “accepted vulnerability to another’s possible but not expected ill will” (p. 235). With this definition, she brings two central concepts into the literature of trust: vulnerability and goodwill. In order to trust, the trustor needs to make themselves vulnerable. As the risk-assessment views highlight, trusting always involves risk. We expect the trustee to behave in some way, but we are never fully certain that they will actually do so; we cannot be, otherwise, there would be no need for trust. The question is: what makes us take the risk to trust? For Baier, the answer lies in the goodwill of the trustee. When we trust a person, we expect that person to care about us and to be willing to behave accordingly. Trustors trust trustees who seem to care about what the trustor cares about.

According to Jones’ (1996) view, trust is an *affective attitude*. Following the wake of Baier, Jones proposes the combination of goodwill and competence as the main traits that a trustee needs to exhibit in order to be perceived as trustworthy (and therefore be trusted). That is, the trustee normally acts as the trustor expects them because the trustee knows they are being counted on. On their part, the trustor trusts because they think that the trustee cares about the trustor and aims not to let the trustor down. The trustor has the positive expectation that such an aim will move the trustee to act as expected (Jones 1996, p. 8). Jones’ account assumes that there is an emotional link between the trustor and the trustee that influences both of them to trust and to act as expected respectively.

Nguyen (2019), by contrast, does not put the focus on emotion. For him, trust is a default attitude according to which we do not call things into question (pp. 9-10). As he puts it, trusting is relying on something without deliberation. According to this idea, the trustor trusts the trusted entity by making use of their help with no second thoughts, simply assuming that the trustee won’t let them down. Nguyen’s approach differs from Baier’s and Jones’ because it includes non-agential objects (for example, trust in the ground we walk on). However, we can still find in Nguyen the idea that trust is an attitude of the trustor; particularly, a non-questioning one. Defining trust as a *non-questioning attitude* means that trust is a default instinct that arises when there is a need to rely on someone or something. Thus, distrust is the attitude that requires explicit motives (we trust someone unless they give us reasons not to do so).

What I call attitudinal approaches have the common denominator of charac-

---

<sup>16</sup>Several other authors share this view, such as Faulkner (2015), O’Neill (2018), and McLeod (2021).

terizing trust as an optimistic mental position by the trustor about the trustee. According to these accounts, trust is based on the trustor’s perception of the trustee and how such a perception makes the trustor feel. The focus is put on the trustor’s positive expectations towards the trustee, and how such expectations shape the relationship between them. The main advantage of attitudinal approaches over risk-assessment ones is that the former address the trustor’s relationship with the trustee as a key component of trust. Most of these accounts are motives-based<sup>17</sup> since they put focus on whether the trustor cares about the trustee. Such care constitutes the main reason for the trustee to behave as the trustor expects. Addressing the trustee’s motives to act –which typically have to do with the trustee caring about the trustor– highlights the relational character of trust. This is crucial for understanding trust among people. Even when the trustee is not an person but an institution or an artificial intelligent system, addressing the relationship between trustor and trustee is important. I proceed to illustrate this with an example.

Let us go back to case A from section 2.4.1. Alice trusts her doctor’s diagnosis, even though she does not have a personal relationship with her doctor. However, a *relationship* exists between Alice and the medical system, based on her previous experiences with different hospitals, treatments and professionals. Imagine that a different patient, Adam, is also diagnosed by Alice’s doctor under circumstances analogous to Alice’s. His chances of receiving a correct diagnosis are identical to Alice’s. But Adam and Alice are different trustors with different experiences, and therefore different relationships with the medical system. Adam has been raised in an Amish community, sceptical of modern science, and it is his first time visiting a hospital. Regardless of how likely the doctor is of behaving one way or another, Adam’s trust is partly determined by Adam himself. Thus, the relationship between the trustor and the trustee, even if it is not personal, is relevant for trust. In this regard, attitudinal approaches improve over risk-based approaches. By contrast to the latter, the former take into account the relationship between trustor and trustee and, as the example shows, such a relationship is an important aspect of trust.

The main weakness of attitudinal accounts of trust is that they portray it as subjective, relegating the rational aspect of trust to the background. Attitudes are psychological constructs that refer to the mental and emotional states of individuals (Perloff 2003, p. 87). According to attitudinal accounts, trust attitudes are based on the trustor’s perception of the trustee and the former’s feelings towards the latter. This characterization of trust focuses on its subjective aspect and it does not consider that trust can be the product of a deliberative process. It is not uncommon that the trustor finds it difficult to make their reasons to trust explicit. This is often the case when the relationship with the trustee belongs to the personal domain, as it is the case with friends or family. However, there are different kinds of trustees and,

---

<sup>17</sup>Most of attitudinal accounts are motives-based, even though that is not the case for the particular form of trust that Nguyen (2021) presents in his paper. The strengths that I highlight in this section are more typical of a (big) subset of attitudinal accounts of trust that is sometimes referred as “will-based accounts” (Jones 1999, p. 68). According to this kind of views, under which both Baier (1986) and Jones (1996) fall, the trustee is seen as trustworthy when they are motivated by goodwill.

in occasions, being able to rationalise trust is key in order to place it correctly. Following the discussion from section 2.4.2, that seems to be the case for AI.

Often, no logical reasoning can be applied in order to generalize the trustor's personal experiences. However, when it comes to groups, institutions or AI, the trustor's reasons to trust should be able to be generalized beyond the trustor's subjective experience, or so it seems. Both institutions and AI have a social character and, in order to be successfully implemented, they should be accepted by a majority of the users in a rational way. If there are no rational reasons to trust a trustee, only the trustors with a personal positive experience or drive will do so. It would be rare if that were the case for the majority of the population. Most of the population probably lacks direct personal experience with the trustee at hand.<sup>18</sup> If only a minority trusts an institution, such an institution won't be able to implement the social organization that it aims to implement. Similarly, if only a minority of users trust an AI system due to personal reasons, the system won't be able to be successfully implemented in society. In order to get a wide number of trustors on the same boat, there must be rational reasons that a wide spectrum of people can recognise as valid and be convinced by. In other words, there are cases in which trust can and should be justified. However, attitudinal approaches do not acknowledge this feature of trust.

Describing trust as an attitude is a stance that overlooks the rational aspect of trust and, therefore, it disregards the possibility of justification. If trust was based merely on the trustor's feelings and intuitions about the trustee, a trustor would never be able to offer rational reasons to justify their trust in a trustee. Thus, trust would never be justified or unjustified. However, in everyday life, we encounter examples of trust to which it is possible to attribute rationality, and rationality is in fact crucial for placing trust in trustees who are worthy of it.<sup>19</sup> Because of this need for justification, a successful account of trust should be able to reflect forms of trust that can overcome subjectivity.

### 2.5.3 Trust as a contract

While risk-assessment approaches failed to capture the relational and more intuitive side of trust, attitudinal accounts of trust struggle to reflect forms of trust that can be justified. A possibility to capture the intuitive side of trust while still seeing it as something that can be made explicit is seeing trust as a contract.<sup>20</sup>

In this section, I explore views that use the metaphor of the contract to conceptualise trusting relationships. A contract is an agreement between at least two parties that determines and governs the rights and obligations between them. Contracts make expectations explicit and determine what a party can legitimately expect from

---

<sup>18</sup>For example, let us say that we talk about trust in police. A wide sector of the civil population says to trust the police, even though most of people have never directly interacted with a police officer.

<sup>19</sup>An example would be trust in the scientific community. A wide part of society would say to trust this community, accepting as true their conclusions without necessarily verifying them by themselves. However, it is not a matter of blind trust. There are reasons to believe that the scientific community states the truth, such as that their reasoning process is available to be verified or that, historically, it has offered knowledge and theories that to this moment are accepted as valid.

<sup>20</sup>For a formalization of contractual trust applied to artificial trustees specifically, see [jacobi2021](#).

the other one. According to Bellia (2002), “promises and contracts share the same instrumental value, they facilitate and encourage reliance and promote cooperation and mutual trust” (pp. 32-33). In this sense, trust can be seen as a non-enforceable implicit contract. The trustor expects certain behaviour from the trustee and, aware of such an expectation, the trustee aims to avoid letting the trustor down.

It is important to notice that the contract metaphor uses the notion of *social contract* rather than *legal contract*. While the latter is legally enforceable, the concept of social contract attempts to capture how social relationships work and why they work that way. By social contract, I refer to the implicit agreement among the members of a society to adhere themselves to rules that benefit them as a group. This way, social contracts explain how and why individuals sacrifice some of their personal freedom to obtain social benefits, such as peaceful cohabitation or state protection. Conceptualising trust as contractual highlights its instrumental value, both for the trustor and the trustee. Trust is valuable because it enhances cooperation: it allows the trustor to delegate something to the trustee without the need to check on them (McLeod 2021). Even in cases in which the trustee has nothing to gain, generally, the trusting relationship benefits them too in terms of social status or meaningful attachments (Kirton 2020).

Charles Fried (2015) states that “in both speech and promising there is *an invitation to the other to trust*, to make himself vulnerable; the liar and the promise-breaker then abuse that trust” (p. 17, emphasis mine). The main focus of Fried’s work is not trust but contracts. So he does not say that trust can be seen as a contract, but rather that contracts require trust. One could argue that if there is trust, a contract is not necessary.<sup>21</sup> Trust itself *is* the contract: there is no need to make explicit the obligations that the trustee has towards the trustor. This is so precisely because the trustor trusts, so they already know (or at least they expect) that the trustee will do their part. However, the expectations and commitment involved in trusting relationships, in contrast with contracts, are not explicit. Then, when trusting relationships fail, there are no necessarily material consequences. This does not mean that breaking someone’s trust has no consequences at all. The consequence of broken trust is a feeling of betrayal (Baier 1986, p. 235). Following this parallel, it is possible to conceive trust as some sort of social contract, in which the trustee commits themselves to do what the trustor expects. If the trustee fails to do so, the contract would be broken and so would trust.

In her seminal work, Baier (1986, p. 251) expresses that contracts are *a kind* of trust, in particular, an extreme one. At the other extreme would be the kind of trust children have in their parents or a religious person in God. While the promisor expects the promisee to stick to the contract, the risk of failure is present enough for the promisor to have the need to formalise the agreement and the consequences in case it is broken. Conversely, a child trusts their parents enough to not even consider the possibility of failure.

Finally, I would like to include Hawley’s (2014) theory among the views that characterize trust as a contract. She presents trust as a commitment. Although

---

<sup>21</sup>Of course, this is rather a philosophical conjecture than a legal claim. Explicit contracts in law have a different status than implicit social contracts as conceived in philosophy. When I talk about contracts I refer to the latter sense, even though Fried and Bellia focus on the former.

commitment can be seen as a kind of contract, the notion has further connotations. According to Hawley “to trust someone to do something is to believe that she has a commitment to doing it, and to rely upon her to meet that commitment” (p. 10). For Hawley, commitment is precisely what distinguishes trust from reliance. While a trustor can rely on a trustee (as Kant’s neighbours did to set their clocks), if the trustee did not accept a commitment, the trustor cannot feel betrayed if the trustee fails to do what they expect. Hawley’s commitment approach is non-motives-based. According to her, as long as the commitment is fulfilled, the trustor will see the trustee as worthy of trust, regardless of what motivated the trustee to fulfill such a commitment. This aspect of Hawley’s account has faced criticism, mostly based on the argument the trustor can feel betrayed even if the trustee sticks to their commitments. Andrew Kirton (2020) illustrates this idea with the example of a sinking boat full of navy men and enlisted men. Navy men have the commitment to save the officers first. However, in Kirton’s example, officers are relatively safe on lifeboats while the enlisted men were struggling in the water. If the navy men adhere to their commitment and fail to prioritise the enlisted men, the latter would rightfully feel betrayed. This example points out that trust and commitment can come apart, revealing Hawley’s theory as incomplete.

What I find most valuable about theories on trust that describe it as some sort of contract is that they introduce the idea of trust as a promise or invitation.<sup>22</sup> A person can invite another to trust them with their behaviour, or explicitly ask them to do so. Analogously, institutions promise to the public to function correctly. For example, institutions such as the police or the government promise social order or the efficient management of a nation, respectively. They invite us to trust them, to participate in society following the rules they set. From such an invitation, it follows the social expectation.

In addition, seeing trust as a contract makes it easier to rationalise it. The reason is that, from a contractual view, it is easier to make the terms of the trusting relationship explicit. Outlining the trustor’s expectations explicitly, makes enables the analysis of whether such expectations are reasonable or not. This is desirable for humans-AI relationships, as well as for other kind of trustees such as institutions.

Regarding artificial trustees, the downside of contractual views on trust is that they assume a certain degree of agency of the trustee. Non-agential promisors cannot really promise anything, and neither they can acknowledge the responsibility of a broken promise. This idea would affect also other kinds of trustees such as children. Normally, it is considered that children (and other people with limited agency) cannot hold commitments due to their lack of self-awareness and understanding of the concept. However, parents still trust their children to behave according to the education they have received, for example.

Lastly, as I already remarked, Hawley’s theory in particular does not address the motives of the trustee to act. I will argue that in trusting relationships, the trustor expects the trustee not only to act as the trustor expects but also to do it motivated by reasons deemed appropriate by the trustor. The trustee’s motivation to carry on commitments matters to the trustor as they reinforce (or weaken) the degree of

---

<sup>22</sup>This concept appears as well in Keren (2014, pp. 12–13)

commitment. Going back to case C, consider the scenario in which Christian carries on with the commitment of taking care of Claire’s cat, but only because she asked and he has a high sense of duty. Most likely, Claire would trust Christian more if what motivates him to take care of her cat is a genuine feeling of fondness towards the cat. A non-motives-based theory of trust as the one Hawley proposes fails to capture this aspect. Non-motives-based theories of trust do not fit well with the ‘trust in AI’ either. Case B is an example of it. The ‘commitment’ of the recruiting system (if we accepted that it could commit to something in the first place) would be to find the best candidate. However, the system is only trusted if it is able to provide adequate reasons for its decision, as analogous to human motives. No one would trust a system that provides technically successful results out of luck, even if it happens to be lucky every time.

### 2.5.4 Recapitulation

In section 2.5, I have reviewed the main accounts of interpersonal trust. The goal was to extract the main characteristics of trust that appear in the existing literature and to contrast them with the features of ‘trust in AI’ proposed in section 2.4.2. Serve table 2.1 as a summary of this contrast exercise.

	Strengths	Weaknesses
<b>Risk-Assessment</b>	<ul style="list-style-type: none"> <li>○ Rationalise trust (potential intersubjectivity).</li> <li>○ Motives-based.*</li> </ul>	<ul style="list-style-type: none"> <li>○ Trust = Reliance</li> <li>○ Oversees the trustor-trustee relationship.</li> </ul>
<b>Attitude</b>	<ul style="list-style-type: none"> <li>○ Relational trust.</li> <li>○ Motives-based (trust ≠ reliance).</li> </ul>	<ul style="list-style-type: none"> <li>○ Highly subjective.</li> </ul>
<b>Contract</b>	<ul style="list-style-type: none"> <li>○ Rationalise trust (potential intersubjectivity).</li> <li>○ Relational trust.</li> <li>○ Trust as an invitation.</li> </ul>	<ul style="list-style-type: none"> <li>○ Trust requires agency.</li> <li>○ Non-motives-based (Hawley).</li> </ul>

(\*) Address whether the trustee is motivated, but not the particular motivation.

Table 2.1: Summary of the main strengths and weaknesses of different kinds of (interpersonal) trust accounts.

In section 2.4.3, I proposed six features of trust, based on the common points between the posit concept of ‘trust in AI’ and the intuitive notion of trust in humans. Those features can be briefly summarized as follows:

- (1) Trust is a relationship in which the trustor has positive expectations towards the trustee.
- (2) Trusting relationships are built on the (potential) need or wish of delegation.
- (3) Trust entails the risk of the trustee not behaving as the trustor wants them to.
- (4) Trust can (and should) be *motives-based* (which makes it distinct from reliance).
- (5) Trust can (and should) be grounded in *rational reasons*.
- (6) Trust can (and should) be *intersubjective*.

In this section, I have looked for parallels between the characterization above and existing accounts of interpersonal trust. Features (1)-(3) seem to be unproblematic and constitute common points between trust in humans and “trust in AI”. Points (4)-(6) can be found spread across the strengths column of table 2.1 (key words in italics). Thus, *some* accounts on interpersonal trust reflect *some* of the features that, I argue, are shared between trust in humans and ‘trust in AI’. However, there is no view that covers all my posit features at the same time. From this, I conclude that the analogy between trust in humans and ‘trust in AI’ is legit,<sup>23</sup> but none of the existing accounts of trust is able to incorporate the analogy on its own.

This gap in the literature does not come as a surprise. After all, the accounts that I have analysed focus on interpersonal trust and their the aim was never to characterise trust in non-human entities. Against this backdrop, arises the idea to compare trust in AI to institutional trust, rather than to interpersonal accounts. I proceed to briefly explore this possibility.

---

## 2.6 Institutional Trust

So far, I have presented trust as a relationship between two individual parties, using interpersonal trust as the paradigm. However, along the last section, there were recurrent examples in which trust was directed towards groups of people such as communities or institutions. In natural language, it is common to talk about trust in groups.<sup>24</sup> We talk about trust in the medical system or in the police, for example. However, it is debatable whether the interpersonal trust theories can successfully describe cases in which the trustee is a group of people rather than a person. On the one hand, I have shown that trusting someone is related to the motivation of the

<sup>23</sup>Therefore, from now on, I refer to the concept of *trust in AI* without quotation marks.

<sup>24</sup>A note on terminology. Much has been written about groups, institutions and organizations, and different authors (Mathiesen 2006; Fricker 2010; List and Pettit 2011) refer to different things when using those labels. I do not aim to go deep into this terminology discussion here and, for the most part, I will be using ‘group’ and ‘institution’ as interchangeable terms. I consider ‘group’ to be a wider category since every institution constitutes a group of people, but not every group can be considered an institution. When talking about institutional trust, I refer to trust in institutions, even though most of what I say can be applied to groups in general.

trustee; it is hard to defend that an abstract group of people has that kind of agency. On the other hand, the features (1)-(3) from the previous section could be used unproblematically to describe some relationships between individuals and institutions. Often, people have positive expectations from institutions. Such expectations lead those people to delegate something to such institutions, in case they need or wish to do so. Such a delegation always entails the risk of the institution not delivering what it promises. So far, so good. So let me delve into the literature that treats institutional trust in particular.

First and foremost, I find it necessary to clarify what I mean by institutional trust since trust in institutions can be understood in different (and interrelated) ways. As Bachmann (2020) points out, institutional trust can be understood as the trust that a person places in another one in the background of an institution, as the trust that a person places in institutional arrangements or as the trust that a person places in another one who represents the institution. In this section, I mainly refer to the latter meaning, even though it is hard to draw a clear line between the three of them.

Following Lahno (2001), an institution is a particular kind of group characterised for presenting certain behavioural patterns. Those patterns are reinforced within the group and state normative expectations (p. 38). According to him, institutional trust differs from interpersonal trust whilst the former is directed not only towards the members of the institution but towards its rules.

At the beginning of the section, I stated that interpersonal trust is the paradigm of trust. While this claim is relatively uncontroversial, there are authors such as Pettit (1995) who point out the influence of institutional design on trust among individuals. Even though trust typically arises between two parties, facts such as the social climate definitively influence how trusting relationships take place, if they do. This line of thought expands the domain of trust far beyond individual trustors and trustees. Furthermore, it allows the conceptual connection with socio-technical aspect of AI. In cases A and B, I used as examples of artificial trustees a medical diagnostic system and a recruiting system. Even though the examples focus on the characteristics and performance of these particular systems, they can be used to show how the perception of the users is not only determined by the particular system but by its role in a bigger network. As we saw when Adam was introduced in example A, distrust is often not only directed towards a particular AIs but towards the whole socio-technical system that it is part of. The same can be said about trust, and in this sense, trust in AI and trust in institutions are tightly interrelated.

Part of the literature addresses the topic of testimonial trust from an institutional perspective (Medina 2020). By testimonial trust, I refer to the acceptance of others' testimony as true. A group can collectively act or provide testimony, meaning that those actions or claims entail collective responsibility, rather than any of the individual members of the group being entirely responsible (Anderson 2012; Fricker 2010). For example, imagine that the police say, via a spoke person S, that suspect A is responsible for certain terrorist attacks and that they have already found and imprisoned this person. Later in time, A is found innocent and a second suspect B is held responsible for the attacks. The spoke person S is under no circumstance responsible for providing false testimony since they were talking in the name of the

police as an institution. Citizens may distrust and recriminate the police after the incident, but it would be rare to blame S in particular and to label them as untrustworthy (in any case, S would be untrustworthy whilst member of the police, no more no less than any other member).

It seems like trusting a collective as such means something different than trusting the set of all of its members. However, it is unclear how to grasp an attitude held by one person and directed to multiple. Furthermore, it is not evident how exactly the relationship trustor-institution works since it is highly asymmetrical.<sup>25</sup> Then, does it make sense to talk about institutional trust? According to Hawley (2017), not really. In her own words, “there is a type of trustworthiness which individuals but not groups can exemplify” (p. 247). Thus, Hawley argues that it is possible to rely on groups but not to trust them (the distinction between trust and reliance applies to the interpersonal domain only). Following a different argumentation, Luhmann (1980) arrives at a similar conclusion. According to him, trust is essentially cooperation. This view makes sense when there is a small number of parties involved, but in the case of institutions (or in Luhmann’s terminology, “systems”), we can only talk about confidence, but not trust (p. 102). Interestingly, there is a parallel to be drawn between this kind of claims and the ones made by authors such as Nickel, Franssen, and Kroes (2010) or Ryan (2020) regarding artificial trustees. Like Hawley, they argue that AI can only be relied on, but not trusted.

Leaving Hawley’s and Luhmann’s criticisms above aside, intuitively, the concept of trust in institutions refers to a rather rational kind of trust. In contrast to interpersonal trust, trust in institutions is rarely based on feelings or intuitions. On the contrary, and echoing Lahno (2001), trusting an institution means trusting its rules and structures. This kind of trust tends to adhere to logical reasoning to a greater extent than trust in people, with whom the trustor may have some emotional bond. Whilst trust is grounded on rational reasons, it is possible to open intersubjective discussion. This is often the case with public institutions, in which social trust is placed following the public opinion that generates from the social debate.<sup>26</sup> Thus, each trustor is largely influenced by other trustors since the object of trust can be discussed and collectively justified.

In this section, I have shown that trust in institutions is related to the posit concept of trust in AI whilst the former is often based on rational reasons and allows intersubjectivity. This way, a significant parallel can be established between the two, sharing most of the features that I disclosed in section 2.4.2. However, as Hawley and Luhmann point out, the fourth characteristic poses a problem: both in institutional trust and trust in AI, it arises the difficulty of distinguishing trust from reliance. Whether this should normatively be the case for institutions is a rather unexplored topic that I won’t address here. Regarding to AI, as I said, human trustors generally expect that an artificial intelligent system offers something more

---

<sup>25</sup>I am not talking here about asymmetric power or responsibility, which is always the case in any kind of trust. What I mean is that while an individual can hold an attitude towards a collective –whatever that means–, how does a collective reciprocate such an attitude? Someone needs to feel some way or take some action, but institutional trust precisely highlights how an institution goes beyond the set of its members.

<sup>26</sup>For an interesting discussion on the value of social trust, see O’Neill (2002).

than accuracy. Neither in the literature of interpersonal nor institutional trust I have found a motives-based theory that successfully distinguishes trust and reliance, while satisfying conditions (1), (2), (3), (5) and (6) at the same time. Therefore, my quest for a theory of trust that can capture trust in people, institutions and AI is so far inconclusive.

Nevertheless, the parallels drawn in sections 2.5 and 2.6 are far from being fruitless. Even though I have not found a existing theory that fits the conditions from section 2.4.2, the literature offers significant common places for trust in humans, in institutions and in AI (see table 2.1). Because of this, I consider that the concept of trust in AI is a legitimate one, even though it differs from other forms of trust that the literature has greater explored to this day. Arguing that trust in institutions or trust in AI are not ‘real’ forms of trust –but reliance– creates a conceptual rupture in the notion of trust. Such a rupture between interpersonal trust and any other of its forms leads to more confusion than elucidation. I agree with the main body of literature that interpersonal trust is the paradigmatic case for trust. However, that does not mean that trust is exclusively a one-on-one human relationship. People commonly use the term to refer to other kinds of trustees such as institutions (Lahno 2001), artificial intelligent systems (Coeckelbergh 2012), the law (Fox-Decent 2005) or oneself (Foley 2001). Arguing that these cases are not ‘real trust’ cases would be going against the natural use of the term. Moreover, addressing other kinds of trustees as possible objects of trust does not corrupt the term ‘trust’, but rather helps to further understand it in its full richness.

I have provided an review of the most popular approaches to both interpersonal and institutional trust. The aim of such a review was to highlight the strengths of each of these approaches while pointing out its problems. When one tries to transpose existing theories of trust to the AI domain, the problems that could already be spotted in the trust literature only get deeper. Still, trust in AI is a highly relevant concept for our society that is worth exploring. Given the common places discussed in this chapter, previous characterizations of trust come in useful to describe trust in AI, despite the gaps in the literature. Understanding what it means to trust AI has become an urgent endeavour in order to clarify if such trust is possible (and desirable) and what it entails in case it is. In the next section, I will revisit the tentative characterization of trust in AI that I presented in section 2.4.2. Drawing on the literature on interpersonal and institutional trust, I will clarify the discrepancies between the accounts discussed in the previous sections and the concept of trust in AI.

---

## 2.7 Revisiting trust in AI

In section 2.4.2, I outlined six key features of trust in AI: the trustor’s positive expectations towards the trustee, the (potential) need or desire for delegation, the inherent risk of the trustee failing to meet the trustor’s expectations, the alignment of motives between trustor and trustee, the grounding of trust in rational reasons, and its intersubjective nature.

In sections 2.5 and 2.6, I provided an overview of the most influential accounts of both interpersonal and institutional trust, aiming to identify parallels with my tentative description of trust in AI and the features mentioned above. I concluded that such parallels do indeed exist, supporting the argument that trust in AI is a legitimate concept. However, these features of trust manifest differently when the trustee is a person compared to when it is an AI system. Before deepening my investigation into what it means to trust AI, there are two points of mismatch worth revisiting:

1. The use of the term ‘motives’ when discussing AI systems.
2. What constitutes a ‘rational reason’ to trust an artificial trustee, as opposed to a human one.

### 2.7.1 Motives

It is relatively uncontroversial to define motives as the combination of desire and belief that drives an agent’s actions (Davidson 1963, p. 687). Intuitively, motives are exclusively human, since AI systems can neither hold beliefs nor have desires. This is a common critique against the concept of trust in AI (Nickel, Franssen, and Kroes 2010, p. 442; Ryan 2020, p. 10). However, when it comes to AI, I adopt a broader definition of motives, considering them as the criteria that favours a course of action. I go back to case C from section 2.4.3 as my go-to example. Let us analyse Christian’s motives to take care of Claire’s cat. Christian is willing to care for Claire’s cat because he likes cats and he believes that the opportunity to take care of Claire’s cat will bring joy to both of them (and hopefully the cat). I call this motivation  $m$ , being  $m$  the reason for Christian to do  $\delta$  (take care of Claire’s cat). Christian’s motivation  $m$  acts as their reason to do  $\delta$  because the expectation of future joy favours Christian to do  $\delta$ . Analogously, there are AI systems that operate following certain criteria that favours them to provide certain outcomes. For example, let us consider a NN designed for brain tumour detection and classification, such as the one described in case A (from section 2.4.1). On some test data, the NN correctly identifies tumours in 99% of the MRI scans it processes, affirming its accuracy. After some time, it is discovered that the NN detects tumours not by focusing on the pixels of the MRI scan in which the tumour can be found, but based on some other non-causal correlation factor. The process that the system follows to achieve results is crucial to determine whether it is worthy of trust or not. In a wide sense, I understand the criteria used by the system to provide its outcome as the system’s ‘motive’ to act. Thus, analogously to how human doctors can be motivated to treat their patients by their professionalism, altruism, scientific curiosity, etc., an AI diagnostic system can be ‘motivated’ to aim for accuracy by its design. In this case, the AI system operates based on the criterion that a certain pattern of pixels typically corresponds to the presence of a tumour. Not only the system’s accuracy is important, but also why the system is accurate. In this sense, I establish the parallel between trust in AI and trust in humans: in both cases, trust should be motives-based. That means that the motives that lead a system to perform as it does play a central role when determining whether the system deserves human trust.

One could ask whether this way of attributing motives shifts the motivation back to the system’s designers, thus leading to the conclusion that only human designers, not the systems themselves, can be trusted. When it comes to this point, I find it helpful to shift the analogy to institutional trust, rather than interpersonal trust. In section 2.6, I explained how in cases of institutional trust, the members of an institution represent such an institution with their individual actions and behaviours. Thus, institutional motives are represented by individuals through individual actions, which transcend the individuals themselves. Similarly, the designers of AI systems encode their motives into the systems they develop. Such systems will execute these motives based on their programmed algorithms and operational settings. If one understands the relationship between AI systems, human designers and their motives in this manner, the following question arises: Why talk about trust in AI specifically? Aren’t we dealing with a case of institutional trust? What distinguishes AI systems like ChatGPT from products like Bosch washing machines? As I noted, there is an analogy to be drawn between trust in AI and institutional trust, but there are some fundamental differences that set them apart. Such differences arise from the special kind of technology that AI is, different from other products that are produced by institutions. Consider the example of a washing machine produced by Bosch, which is developed by a team of engineers representing the company. Similarly, ChatGPT is developed by a team at OpenAI. The main difference between ChatGPT and a washing machine is that the latter does not develop its own criteria to ‘decide’ which course of action to take. A washing machine operates on fixed, deterministic programs without the capacity to modify its operations based on external data. In contrast, NNs like the one powering ChatGPT dynamically adjust their weights during the training process. Even if this process is predetermined, it is not possible for the designers to anticipate every possible system’s outcome, thus giving the AI a semblance of developing its own criteria for action.<sup>27</sup>

### 2.7.2 Agency

As the reader can likely anticipate, the distinction between AI having motives and seeming *as if* AI had motives is blurry. This brings me back to section 2.3.2, where I established that AI can be considered quasi-agents. It is worth expanding on that idea now, to explain why I refer to quasi-agency rather than full agency, and to address whether someone or something can truly be the object of trust without being an agent.

Agency is traditionally understood as the capacity to act intentionally and for reasons. In her seminal work *Intention*, Anscombe (1957) emphasizes that intentional actions are defined by the agent’s knowledge of what they are doing and why they are doing it. For Anscombe, an agent’s reasons are not merely causes but explanations of the agent’s purpose, driven by practical reasoning (Anscombe 1957, pp. 9–12). Building on this foundation, Davidson (1963) argues that agency involves the ability to perform intentional actions, which are driven by the agent’s beliefs and desires. The combination of belief and desire is labelled by Davidson as ‘reason’, but can also

---

<sup>27</sup>As I pointed out in section 2.3.2, AI systems can be considered quasi-agents due to being perceived *as if* they were agents by the humans who interact with them.

be understood as intention or motive. Thus, in this framework, motives derive from mental states, serving both as the causes of actions and as justifications for them (Davidson [1963], pp. 686–687). According to both Anscombe and Davidson, acting intentionally means acting for reasons that explain and motivate one’s behavior. In this standard conception of action, the agent’s actions are motivated by mental states such as beliefs, desires, or a combination of both, which, in early Davidson’s view, correspond to motives.

A common critique of AI’s agency is based on the fact that AI systems lack mental states and, consequently, motives, which excludes them from being considered agents. However, as I discussed in the previous section, if we adopt a broader notion of motives, where they are seen as criteria that favor a course of action, it could be argued that AI systems do have motives. The question then becomes: does this broader understanding of motives justify referring to AI as agents? To answer this, I need to explore the concept of agency further.

The standard conception of agency holds that an individual is an agent if they are capable of possessing certain mental states that guide their actions. This view reflects an architectural approach to agency, where individuals are either agentic or not based on their inherent attributes. Such attributes typically include the capacity to hold certain mental states, but some authors extend beyond the immediacy of mental states such as desire or belief to wider preconditions such as rationality (Anscombe [1957], p. 9), autonomy (Frankfurt [1971], pp. 6–7), accountability (Strawson [1962], pp. 11–12), the resilience required to adhere to one’s plans (Mele [1992], p. 125), or the purposefulness of action (O’Brien [2022], pp. 109–110). Within these kind of architectural approaches, it is difficult to justify classifying AI systems as agents, as they lack many of the aforementioned attributes traditionally associated with agency.<sup>28</sup>

However, some authors conceptualize agency in behavioural terms rather than architectural ones. Instead of classifying individuals as agents or non-agents, the focus shifts to determining which behaviours are agent-like; identifying which individuals exhibit such behaviour becomes a secondary question. A well-known example of this kind of strategy is Daniel Dennett’s ([1988] *intentional stance*). In the book of the same name, Dennett describes what he calls the intentional stance. This strategy consists of attributing mental states, such as beliefs and desires, to a system *as if* it were an agent, aiming to predict its behaviour. For Dennett, this is an instrumentalist strategy; the aim is not to figure out the system’s true nature, but to predict its behaviour. In this way, Dennett proposes the intentional stance as a means of simplifying the workings of complex systems, making it easier for human agents to predict the future actions of such systems. In contrast to the intentional stance, Dennett proposes two other ways to predict a system’s behaviour: the physical and design stances (Dennett [1988], pp. 16–17). The physical stance allows the prediction of behaviour based on the laws of physics and the physical constitution of the system. The design stance allows the prediction of behaviour based on how the system is designed to function. For example, one could aim to predict when a child will grab

---

<sup>28</sup>Whether AI systems exhibit rationality is debatable and depends on how we define rationality. Other attributes typically linked to agency, such as autonomy, seem less problematic when attributed to non-human entities like AI systems.

a cookie by adopting the physical stance, focusing on the child's biological processes. This would involve examining how long it takes for their last meal to be digested, when their body signals hunger, and how this drives them to seek food. By observing the mechanics of their body –how hunger triggers action and their muscles reach for the nearest snack– it is possible to predict that the child will grab the cookie. Alternatively, by taking the intentional stance, one could predict the child's behaviour by attributing mental states, such as the desire for the cookie and the belief that it will satisfy their hunger. Without needing to know the body's internal workings, it is inferred that the child will grab the cookie because they want it and believe it will give them pleasure. While both strategies lead to the prediction of when the child will grab the cookie, the intentional stance offers a convenient shortcut in the reasoning needed to make such a prediction.

The main advantage of Dennett's strategy is that it avoids the essentialist question of whether a system is an agent, thus bypassing the difficulty of defining what constitutes agency. The intentional stance can be applied to a wide range of complex systems, regardless of whether they truly are agents. The strategy remains effective as long as the system's behaviour can be reliably predicted by attributing agential attributes to it, such as beliefs, desires, or intentions. Dennett argues that this approach can be applied to computers and AI systems, noting that, in his own words, it would work just as well for predicting the behaviour of a chess-playing machine as for a homunculus hidden inside one (Dennett [1988], pp. 74–75). In the case of AI systems, understanding them as intentional agents in Dennett's sense makes it easier to manage what to expect from them as a user. By treating AI systems as intentional agents, we can predict and interact with them more effectively, even if they do not possess true consciousness or intentions. Using the intentional stance, we predict AI behavior by attributing to it certain motives, similar to how we understand human actions. By considering AI systems as quasi-agents with motives, I align my view with Dennett's approach, simplifying the prediction of their behaviour.

Returning to the topic of trust, Dennett's intentional stance provides a useful framework for understanding how we might extend trust to entities that are not traditionally considered agents. While Dennett's approach allows us to predict behaviour without requiring true agency, the question remains: does trust itself necessarily require agency? Traditionally, trust has been thought to require agency, since trust typically involves expectations about the motives, intentions, and actions of another. Trusting someone usually implies that the person or entity being trusted can make decisions based on those motives and intentions, which is what we associate with agency. However, I have already clarified that my understanding of motives is wider than the mainstream use of it, which makes it applicable to non-human entities. Nevertheless, this does not mean that any kind of entity can exhibit motives or be the object of trust. The question now is: does trust require agency? I will argue that it does not. The reason is that in a trusting relationship, the trustor never has direct access to the trustor's motives.<sup>29</sup> The trustor trusts based on their interactions with the trustee. That is, based on the motives that the trustor attributes to the trustee. The intentional stance supports this by offering a method for interpreting

---

<sup>29</sup>Or their mental states, if one wanted to stick to Davidson's view of motives/reasons/intentions as being a combination of states such as beliefs and desires.

AI behavior as goal-oriented or intentional, allowing for the establishment of expectations. Furthermore, it is worth noting that the intentional stance is something we, as humans, bring to the system. We choose to adopt the stance because it simplifies our understanding. This aligns with my view of AI systems as quasi-agents, as presented in section 2.3.2. If humans consistently interact with AI as if they were agents –using the intentional stance– then the AI can fulfill a similar role to human agents in trusting relationships, even though it does not possess true agency.

In sum, the users interact with devices that act as if they had motives, and based on such a perception, a human-AI relationship is shaped. I label this motives-based relationship as trust. Because this kind of trust is analogous, but distinctively different, from trust in people or institutions, the notion of ‘trust in AI’ becomes necessary.

### 2.7.3 Rational reasons

When discussing the features that characterised trust, I argued that trust should be grounded on rational reasons.<sup>30</sup> In the case of human trustees, the trustor’s reasons to trust the trustee often refer to the trustee’s character, traits or the personal relationship that the trustor has with them. None of these reasons seem to apply to artificial trustees, with whom humans do not form personal bonds.<sup>31</sup> However, note that when describing trust in AI, I talked about *rational* reasons, not just reasons of any kind. By rational reasons, I mean those that can be legitimately used to justify trust and, therefore, have the potential to be agreed on by others than the trustor. This kind of reasons should constitute a solid ground (by themselves, or in combination with further reasons) to believe that the trustee is competent and that there is a motive alignment between the trustor and the trustee. These kind of reasons can be found in cases of interpersonal or institutional trust, or trust in AI.

For both human and artificial trustees, rational reasons to believe in their competence function similarly. A rational reason to trust that a trustee is capable of a certain task might be based on observing the trustee successfully performing that task in the past. This reasoning applies equally to both human and artificial trustees: just as I might trust someone to care for my cat because I have seen them take care of other cats, I might trust ChatGPT to translate a text because I have seen it effectively used by others for the same purpose.

When it comes to justifying motives alignment, human and artificial trustees differ notably. Trustors typically infer a human trustee’s motives based on their personality or social cues. For instance, in case C,<sup>32</sup> Claire believed that Christian’s love for cats was a motive for him to take care of hers. This type of reasoning does not translate directly to artificial trustees. However, trustors can access the motives of AI systems (understood as outlined in the previous section) when the system is

---

<sup>30</sup>For details, see section 2.4.2, under the heading “(5) Rational reasons”.

<sup>31</sup>However, I do not rule out the possibility of this circumstance changing in the future. Given the rapid pace of AI development today, it is plausible to think that, in the near future, enhanced by embodiment of AI in robots, human trustors may begin to attribute typically human qualities to AI trustees. This would be similar to how we already treat many of these systems as if they were agents, even though we know they are not (see section 2.3.2).

<sup>32</sup>See section 2.4.2.

explainable, as illustrated in case B.<sup>33</sup> This topic warrants further discussion, but for now, I leave it at that. I will revisit this issue in later chapters when I have developed my ideas on trust further and I am better equipped to argue about how trust is built and how human trustors perceive artificial trustees.

---

<sup>33</sup>See section 2.4.1.



# 3

---

## *Trust as a belief-based disposition*

---



## 3.1 Introduction

Trust is a rich and multifaceted concept that can occur in a variety of circumstances. I trust my mother, my friends, my boss, the medical system, the judiciary and, sometimes, I even trust myself. There are different kinds of trustees that establish very different kinds of relationships with trustors. However, in ordinary language, we refer to all of those relationships as ‘trust’.

In chapter 2, I considered the notion of ‘trust in AI’ by making explicit the expectations that humans have towards AI in cases in which trust is required. I compared these expectations with what is typically expected from human trustees, both when the trustee is an individual person or an institution. From this comparison, I carved out the similarities between the three kinds of cases (artificial, interpersonal and institutional trustees). I argued that such similarities are strong enough to conclude that AI systems can be legitimately referred to as trustees. In other words, trust is a concept that can be used to describe not only human-human interaction but also human-AI interaction. However, the existing literature on trust has not really explored the latter case.<sup>1</sup> The aim of this chapter is to fill that gap by developing an account of trust that reflects not only cases in which the trustee is human but also the possibility of artificial trustees.

A successful account of trust should be wide enough to clarify how trust works in all its different manifestations. However, it is important that the account is not so wide that it accidentally encompasses related concepts such as reliance, mistaking it for trust. To this end, the new account of trust must overcome the challenges that its predecessors faced. On the basis of the literature review from chapter 2, I highlight three key features of trust that a proper account should be able to capture but that so far have been proven to be problematic:<sup>2</sup>

- (4) Trust is grounded not only on the likelihood of successful delegation taking place. The trustee’s motives to act should matter to the trustor too. Because of this, trust is motives-based, which makes it distinct from reliance.
- (5) Trust is not always irrational. On the contrary, it can (and it should) be based on rational reasons.
- (6) Despite its subjectivity, trust can be intersubjective when based on rational reasons.

In order to develop a theory that meets the above conditions, I propose focusing on the doxastic root of trust; that is, on the fact that trust originates in belief. A key aspect of trust is that it is not directly based on trustworthiness, but on perceived trustworthiness. In other words, the trustor trusts the trustee when the former *believes* that the latter is trustworthy (which may be the actual case or not).

<sup>1</sup>There are three worth-noting exceptions that I am aware of: Ferrario, Loi, and Viganò (2020); Jacovi et al. (2021); Alvarado (2023). Even though they reach different conclusions than the ones I will offer in this thesis, they provide valuable conceptual analyses of ‘trust in AI’.

<sup>2</sup>The list begins at (4) for consistency with chapter 2 (see section 2.4.2).

A trustee is trustworthy when it is reasonable to delegate a task or a role to them. On this basis, trusting a trustee means believing that there are good reasons to delegate something to such a trustee. According to this conception of trust, the trustor's beliefs and their basis are key. Let me explain how pointing at beliefs as the origin of trust sheds light on points (4)-(6).

Trust, as I present it, requires that the trustor believes that the trustee is trustworthy. This has been called a *trust belief* (Keren 2020, p. 109). Under that assumption, focusing on the doxastic aspect of trust will allow me to distinguish it from other concepts such as reliance, which can occur without the presence of trust beliefs.<sup>3</sup> Thus, understanding the role of beliefs in trusting relationships sheds light on point (4): trust is motives-based (see section 1.3), and the trustor's attribution of motives is strongly related to the trustor's beliefs about the trustee. In other words, a trustor  $A$  attributes a motive  $M$  to a trustee  $B$  based on  $A$ 's previous beliefs about  $B$ .  $A$ 's belief that  $B$  is motivated by  $M$  to do a task  $\delta$ <sup>4</sup> triggers  $A$  to trust  $B$ .

In turn,  $A$  believing that  $B$  holds motive  $M$  generates further beliefs in  $A$  (about  $B$ ). For example:

- ( $b_1$ ) Claire believes that Christian has several cats.
- ( $b_2$ ) Claire believes that Christian treats his cats very well.
- ( $b_3$ ) Because of ( $b_1$ ) and ( $b_2$ ), Claire believes that Christian loves cats.
- ( $b_4$ ) Claire believes that ( $b_3$ ) acts as a motive  $M$  for Christian to take good care of her cat if she asked him.
- ( $b_5$ ) Furthermore, believing ( $b_3$ ) makes Claire believe that Christian is a good person.

In contrast, motive attribution is not necessary for reliance to occur since reliance is necessarily tied to delegation. There is no such scenario in which a subject relies on another with no delegation occurring and, therefore, having to consider how they would act in case delegation was required.

The doxastic aspect of trust also explains point (5): the duality between rational and irrational trust. In the same way that beliefs can be either rational or irrational depending on how they relate to each other, so is the case with trust. Trust beliefs can be supported by other beliefs from diverse natures. For example, I could trust my doctor just because they happen to be a close friend. But I could also trust them because I believe that medicine is a successful science, that the medical system works successfully, that my doctor got educated in a reputed institution which abides by such a system, and so on. The network of beliefs that support a trust belief can be more or less coherent. Thus, characterising trust as belief-based sheds light on how

---

<sup>3</sup>Recall the characterization of reliance offered in section 1.2. There, I explained that it is possible to rely on people or objects even without believing that they will most likely succeed at the delegated task (for example, one can rely on a shaky ladder to change a light bulb, in case there are no other means available).

<sup>4</sup>And, additionally, that  $B$  is capable to do  $\delta$ .

that trust can be rational, irrational, or something in between, even though only the former is desirable.

Finally, point (6), which referred to trust's potential intersubjectivity, also finds its explanation in the way the trustor's beliefs are formed. While beliefs are subjective in the sense that they depend on the subject who holds them, they can also be shared by different subjects. This applies to trustors' beliefs on common trustees: different trustors can agree on their beliefs about a trustee, especially when such beliefs have a rational basis (independent of each trustor's personal experience). Trust is rational when the trustor's belief that the trustee is trustworthy is coherent with the trustor's belief system and accepted evidence. When this is the case, trust is not merely subjective and the door to intersubjectivity opens up. The reason is that the availability of rational reasons allows different trustors to compare their beliefs and evaluate them on a rational basis. Ideally, this is how trust should always be, to increase the likelihood of placing trust only in trustworthy trustees.

Characterising trust as belief-based relegates it to the mental realm. However, trust is also closely linked to action. Actual delegation does not need to occur for a trustor to trust, but the hypothetical possibility of delegation must be feasible. There is something lacking in doxastic approaches to trust to fully capture the latter aspect, or so it seems. Because of this, I introduce the concept of disposition. A disposition is a state or condition that, under certain circumstances, triggers further specific behaviour. This way, the term disposition is used to describe states which not only refer to the actual state of the world but to future states that could possibly follow from it. For example, an object is soluble *if* it would dissolve in the event of being submerged in water. Thus, solubility is a disposition. I argue that trust presents a similar structure: a trustor trusting a trustee means that the former would delegate something to the latter *if* the trustor wished or needed to do so.

If we accept the above definition of trust, trust does not only concern the actual world, but possible worlds that could materialize under certain circumstances. For a possible world in which the trustor delegates something to the trustee to materialize, the trustor needs to hold a trust belief in the actual world. This is what I call a *belief-based disposition*: an individual's state in which they would behave in a certain way given certain circumstances, as long as they hold a certain belief. On this basis, I define trust as a belief-based disposition because trust is the state of a trustor who would delegate something to a trustee, as long as the trustor believes that the trustee is trustworthy.

This chapter is devoted to developing the above idea. To do so, the chapter is structured as follows. In section 3.2, I motivate the need for a new account of trust. I recall from chapter 2 the challenges that existing accounts of trust face, and that the theory I propose should overcome. In section 3.3, I develop my view on trust in detail. I propose the concept of belief-based disposition to characterise trust and to shed light on its workings, both for human and artificial trustees. My aim is to revisit the picture I painted in chapter 1 and use the belief-based disposition account to redefine both trust and trustworthiness. By the end of the section, I go back to the challenges pointed out at the beginning of this chapter to check whether my account tackles them better than its predecessors. Finally, in section 3.4, I explore and contest possible objections to my view.

## 3.2 Motivation

In this section, I motivate the need for a new account of trust. First, I will describe how a successful account should be and the challenges it should be able to overcome. Then, I will develop the tensions that follow from those challenges and propose ideas to resolve them.

### 3.2.1 Challenges

Chapter 2 concluded with the idea that trust in AI shared enough features with trust in humans to be considered a legitimate kind of trust. The common features that allowed me to reach such a conclusion were the following:

- (1) Trust is a relationship in which the trustor has positive expectations towards the trustee.
- (2) Trusting relationships are built on the (potential) need or wish of delegation.
- (3) Trust entails the risk of the trustee not behaving as the trustor wants them to.
- (4) Trust can (and should) be motives-based (which makes it distinct from reliance).
- (5) Trust can (and should) be rational.
- (6) Trust can (and should) be intersubjective.

The difficulty lies in capturing all the features above by a single account of trust at the same time. Features (4)-(6) proved to be particularly challenging. The reason is that they are normative rather than descriptive. That is, they point out how trust *should* be and signal the fact that it is indeed possible that trust manifests itself in those ways.

Normativity is the quality that characterizes action-guiding claims. That means that a normative claim, or feature, does not merely state a fact or describe a situation. Normative features point to how something should be, not how it is. Then, normativity always involves some kind of tension between reality and norm. In the case of the features (4)-(6), the following tensions arise:

- (4)' Trust is often based on whether delegation is (likely to be)<sup>5</sup> successful or not, while the trustee's motives to perform the delegated task should be taken into account as well.

---

<sup>5</sup>I include this parenthesis here because, as stated in chapter 1, trust refers to future hypothetical scenarios. Thus, first the trustor trusts, and only then they (may) delegate a task to a trustee. Trust arises before any delegation takes place; otherwise, there would be nothing to trust about. Therefore, trust can not be based on whether the delegation of a trusted task is successful or not. That would be nonsensical since it is not possible that something (trust) is based on a fact (delegation) that has not happened yet. Because of this, when I talk about trust being based on successful delegation, strictly speaking, I mean trust being based on how likely delegation is to be successful. In the rest of the thesis, I will stick to the former phrasing for simplicity and easier reading.

(5)' Trust is often irrational, while it should be rational.

(6)' Trust is often subjective, while it should be intersubjective.

In the following, I develop these tensions and suggest ways in which they could be resolved.

### 3.2.2 Tensions to overcome

#### Successful delegation versus motives to delegate

In chapter 2, I described how risk-assessment theories on trust (Gambetta 1988; Hardin 2002) put the focus on the likelihood of delegation being successful. Contractual views as Hawley's (2014) also considered that trustors trust based on whether their delegation on the trustee is successful, or in her words, whether the trustee keeps their commitment or not. In contrast, most attitudinal views on trust (Baier 1986; Jones 1996) are motives-based, meaning that what motivates the trustee to act is relevant for the trustor. I do agree with the latter theories that trust is usually motives-based, or at least it should be.

Along this thesis, I have defended that trust is associated with *potential* delegation, unlike reliance, which cannot materialize without *actual* delegation (see section 1.6). Trust can be relegated to mere reliance either when the stakes are low or when the trustor lacks of trustees who they actually trust. For example, Königsberg neighbours could rely on Kant to adjust their clocks because there were no serious consequences if the adjustment was not perfect. Taking a more contemporary example, I often rely on Deutsche Bahn to reach my destination simply because I have no other means of transportation. These kind of situations often take place, but they are not ideal. When the trustor delegates to a trustee a normative task –there is a certain way in which things should be–, then reliance does not suffice. In order to trust, the trustor considers not only the likelihood of the trustee behaving in a certain way but also whether they approve of the motives that move the trustee to act that way.

The above applies not only to human trustees but also to artificial ones. Recalling case B from chapter 1.6 for the system to be worthy of trust, finding the best candidate for a job vacancy was not enough. How the system found such candidates played a crucial role. The fact that the system was biased made it untrustworthy, since the system's 'motives' to select a candidate were not considered appropriate by the company. Therefore, both in cases of trust in humans and trust in AI, trust should be motives-based.

Before moving on, let me note a special case of trust. There can be trusting relationships in which the trustee's motives may be irrelevant to the trustor. I am thinking of cases of blind, irrational trust, in which the trustor bases their trust on the trustee's authority. The greatest exponent of this kind of trust is trust in God. The religious person does not evaluate God's motives and whether they are in alignment with their own motives. The figure of God is unquestionable and, therefore, trust in Him is blind. Other examples include the child who blindly trusts their parents

---

<sup>6</sup>Section 2.4.1

(Baier 1986, pp. 241–242) or the cult member who blindly trusts their leader. These examples illustrate the possibility to trust a trustee regardless of what motivates the trustee to act. However, these examples are far from being paradigmatic and, in any case, they do not portray how trust should normatively be. Children should eventually mature, becoming their trust in others more rational and, therefore, better placed. Cult members should not be deprived of their freedom of thought and they should not blindly follow a leader who may or may not act in their best interest. About the religious person, I refrain from commenting. In conclusion, my remark is that to avoid mistrust, trustors should evaluate what motivates trustees to act as they do and whether the former deem such motives appropriate.

### **Irrationality versus rationality**

Very often, trust is not supported by rational reasons. People trust others because of how they feel about them or as a product of their intuition. These are also reasons, although irrational ones.<sup>7</sup> Even though this is common, that is not how trust ought to be. In most day-to-day cases of trust, a certain degree of irrationality seems to be harmless. However, the higher the stakes, the more important it is to place trust correctly. In order to do so, trust should be based on rational reasons.

There is a tension between irrational cases of trust and cases in which trust is warranted and grounded on good justifying reasons. Trusting relationships can adopt many forms, not all of them complying with normative expectations. For example, I trust strangers when I ask for directions, even though I know that it is unjustified to take as true the testimony of a completely random person whom I know nothing about. It is not my aim to deny this form of irrational trust but to highlight that it is not the only one. While trust can be irrational, it can also be rational. Trust ought to be based on rational reasons when the consequences of misplacing trust are unacceptable.

When trust can be supported by rational reasons (based on the trustworthiness of the trustee), it is *appropriate*. By rational reasons, I mean those that abide by logic, so they can be legitimately used to justify trust. Typically, those reasons are the following:<sup>8</sup>

- The trustor’s motivation is aligned with the trustee’s motivation to act as the trustor wants.
- The trustee has the competence to act according to such motivation.
- Therefore, the trustee will most likely succeed at the task or role that the trustor may delegate to the trustee, if the trustor wished or needed to do so.

When the above conditions are satisfied, it is reasonable for a trustor to trust. But, once again, those are not the only reasons potentially enabling trust. For example, some people trust due to prejudices, even when such prejudices are disconnected from the actual motivation or competence of the trustee. Then, the question is: how

<sup>7</sup>For an explicit description of what I mean by ‘rational reasons’, see section 2.4.2

<sup>8</sup>This list has been constructed by generalising the rational reasons to trust an artificial trustee that I provided in section 2.4.2.

is it possible that both rational reasons and irrational intuitions or emotions can both lead to the same kind of relationship, namely trust? This tension is only apparent since what brings trust in a trustor are not rational or irrational propositions (in the form of reasons) per se. The trustor comes to trust when they *believe* in such propositions.<sup>9</sup> An agent can hold both rational or irrational beliefs and, even though only the first case is desirable, epistemologically, holding irrational beliefs poses no problem.

The trustor can believe in a set of propositions that logically lead to the conclusion of some of the propositions listed above. In this case, the trustor's belief is justified and their trust is rational. But as I said, this is not the only mental process which can result in trust. Often, trustors hold irrational beliefs that can only lead to trust abiding by illogical reasoning.<sup>10</sup> Both cases can be explained by characterising trust as belief-based since beliefs can be rational or not. This dissolves the tension between rational and irrational cases of trust.

### Subjectivity versus intersubjectivity

In chapter 2 (sections 2.4.2 and 2.4.3), I briefly explained how trust can be intersubjective and why that should normatively be the case. Let me expand on that idea.

Often, we trust others without being able to justify why. For example, I trust my mother to take care of my cat, even though she does not have a cat herself and has never looked after one before. However, the kind of personal relationship and shared past experiences that we have, led me to trust her with this duty, even if doing so is not exactly rational. I trust her due to personal motives that have little to do with rational reasoning. Most likely, another person would never delegate the care of their cats to my mother, because they do not share with her the experience as I do. My trust in her is subjective.

To be clear, at its core, trust is always subjective. We trust others based on our perception of them and such perception is untransferable to a different trustor. What I believe about a trustee is based on my personal past experiences with them. You may agree with me on trusting the same trustee I do, but if you do, it will be due to your own personal reasons.

Nevertheless, the fact that trust is subjective does not mean that trust can not be rational. There are cases in which the trustor has perfectly rational reasons to trust the trustee. For example, recall case C from chapter 2. I presented different scenarios in which Claire trusted Christian and, in most of them, she had good reasons to do so. Trusting someone to take care of a cat based on their previous experiences with cats is a reasonable thing to do. Trust is reasonable if it is based on rational reasoning, as is the case when inducing Christian's future behaviour from

<sup>9</sup>The idea that trusting a trustee implies believing that such a trustee is trustworthy is relatively controversial. Authors such as Holton (1994, p. 63), Jones (1996, p. 5), and Faulkner (2015, pp. 428–29) disagree on this. I will properly address this and other criticisms at the end of the chapter.

<sup>10</sup>Remember beliefs  $(b_1)$ ,  $(b_5)$  from the example used in section 3.1. In that case, beliefs  $(b_1)$  and  $(b_2)$  lead Claire to rationally believe  $(b_3)$ , which makes trusting Christian rational (since he loves cats). Imagine that instead of  $(b_1)$  and  $(b_2)$ , Claire believed  $(f_1)$  Christian has fish and  $(f_2)$  he treats his fish fairly well. If, from  $(f_1)$ - $(f_2)$ , Claire deduces that Christian loves cats  $(b_3)$  and, therefore, she trusts him to take care of her cat, Claire's trust would be irrational.

his past experiences. When trust is grounded in rational reasons, it has the potential to be shared among different trustors. This is so because rational reasons overcome mere subjectivity, and can be discussed and evaluated on a rational basis. When trust can be supported by this kind of reasons, it becomes intersubjective.

A successful theory of trust should be able to account for cases of trust in which trust is intersubjective. Otherwise, the theory is incomplete. Sometimes, trust is misleadingly characterised as purely irrational (Baier [1986], p. 235; Jones [1996], p. 11; Nguyen [2021], p. 2).<sup>11</sup> But, as we have seen, that is not always the case. Opening the door for intersubjectivity is particularly important in cases involving artificial trustees. When the trustee is an AI system, subjective reasons should not be enough to trust it, especially in high-stakes situations. For example, an AI diagnostic system could not be implemented in a hospital if only a small group of practitioners trusted it, and their trust was based on personal experiences that cannot be generalized. In the field of AI, trust demands a certain grade of objectivity; or at least the possibility of rational evaluation of the reasons to trust. There is a tension between the subjective taste that characterizes some accounts of interpersonal trust and the normative need for intersubjectivity when the trustee is not human. Conceiving trust as potentially intersubjective solves this tension.

### 3.2.3 Resolution: a doxastic account

In section [3.2.1], I listed key features of trust that a successful account of trust should be able to explain. In section [3.2.2], I developed the apparent tensions that some of those features pose and proposed ideas to conceptually dissolve them. In sum, these are the conclusions from the last section discussion:

- Trust is based on belief. To believe that the trustee is trustworthy, the trustor does not only consider how likely it is that the trustee succeeds at the delegated role or task, but also whether the trustee’s motivation to act is in alignment with the trustor’s motives.
- Trust beliefs can be rational or irrational. To lead to trust, rather than mistrust, trust beliefs should be rational.
- Trust beliefs (and the beliefs that support them) have the potential to be (dis)agreed upon by different trustors.

From the above list, it becomes clear that the resolution to the tensions formulated in section [3.2.2] passes through the adoption of a doxastic account of trust since trust originates in belief. In other words, what I propose is to characterise trust as belief-based. By belief-based, I mean that trust is a phenomenon that cannot occur without the trustor holding a trust belief about the trustee. A trust belief is a belief of the sort of “*A* believes that *B* is trustworthy (regarding  $\Delta$ )” or “*A* believes that *B* will do  $\delta$ ”. By trustworthy, I mean that it is reasonable for the trustor to delegate

---

<sup>11</sup>In no way do the authors above state that trusting is irrational. However, they put the focus on purely subjective and emotional aspects of trust. In this sense, their accounts describe an irrational notion of trust; or better put, a notion of trust beyond rationality.

something to the trustee (see definition 1.3 in chapter 1). Thus, holding a trust belief means believing that a trustee  $B$  is trustworthy, that is, believing that  $B$  would most likely succeed at  $\delta$  (so it is reasonable to delegate  $\delta$  to  $B$ ). Here lies a key difference with reliance: while it is possible to rely on someone without necessarily believing that they will most likely succeed at whatever we rely on them with, it is not possible to trust someone without trust belief. Remember the example I set in section 1.2 when describing reliance. There, I used a shaky ladder to illustrate how one can rely on such a ladder to change a light bulb, in case there are no other means available. However, since trust goes beyond delegation, without trust belief there is no trusting relationship (in this case, between me and the ladder). I may be forced to rely on someone, but that does not mean I trust them. Even if both trust and reliance can result in delegation, the reasons to delegate are different, and so is the whole mental process behind each phenomenon.<sup>12</sup>

On the basis of the discussion above, a tentative reformulation of definition 1.2 reads as follows:

**Definition 3.1.** *Trust* is the belief of a trustor  $A$  that a trustee  $B$  is worthy of  $A$ 's trust; or in other words, that  $B$  is trustworthy. On that basis,  $A$  would be willing to delegate a task or role  $\delta$  (belonging to a domain  $\Delta$ ) to  $B$ , in case  $A$  wished or needed to do so.

But there is a part of the above definition missing. For it to be complete, it is necessary to define trustworthiness as well. Based on this chapter discussion and using definition 1.3 as a starting point, trustworthiness could be defined as follows:

**Definition 3.2.** *Trustworthiness* is the relational property of a trustee  $B$  being worthy of a trustor  $A$ 's trust, in some particular domain  $\Delta$ .  $B$  is worthy of  $A$ 's trust (regarding  $\Delta$ ) iff it is reasonable for the trustor to delegate some task or role  $\delta$  (belonging to  $\Delta$ ) to them. It is reasonable for  $A$  to delegate  $\delta$  to  $B$  iff  $A$  has access to rational reasons that lead to the conclusion that  $B$  is likely to perform  $\delta$  successfully, moved by motives deemed appropriate by  $A$ .

With the above definition of trustworthiness in mind, definition 3.1 would come to say that trust is the trustor's belief that it is reasonable to delegate some task or role  $\delta$  to the trustee. In the literature, it is not uncommon to find authors who define trustworthiness based on trust (Baier 1986; O'Neill 2002; McLeod 2021), rather than vice versa. Baier emphasizes that trust only exists in contexts where trustworthiness is assumed (p. 246). However, I consider trust to be a more fundamental concept than trustworthiness because trust is the prerequisite for recognizing or evaluating trustworthiness. Furthermore, I do not consider trustworthiness to be a fixed quality of the trustee. As trust, I have claimed that trustworthiness is a relational property, meaning it arises out of the trusting relationship. Without trust to initiate the relationship, there would be no way to assess trustworthiness (Hawley 2014, pp. 15–16).

A pure doxastic<sup>13</sup> approach –as the one I proposed– solves some of the difficulties

<sup>12</sup>For details on the contrast between trust and reliance, see 1.6.

<sup>13</sup>Term taken from Keren (2020, p. 109). Doxastic accounts of trust are those accounts that characterise trust as the trustor's belief that the trustee is trustworthy (Hardin 2002, p. 10) or the belief that the trustee will perform a certain act (Hieronymi 2008, p. 216).

that existing accounts of trust encounter. Namely, it explains why the trustee's motivation to act should matter to the trustor: the trustor trusts when they believe that the trustee is trustworthy, and they believe that the trustee is trustworthy when the trustee is moved by motives deemed appropriate by the trustor. However, beliefs are not always rational, meaning that they do not always follow from rational reasoning. As one can believe in conspiracy theories, one can trust without being able to justify such trust. If trust equals trust belief, such an equivalence explains the fact that trust is not always rational. So it does that different trustors can potentially agree on the content of their mutual trust beliefs. However, there are still some problems to come. In the next section, I will address the problems that arise from defining trust as trust belief and come back to (tentative) definitions 3.1 and 3.2.

---

## 3.3 Proposal: a belief-based disposition

### 3.3.1 Preliminary problems

I have proposed definition 3.1 as a tentative definition of trust, putting the focus on its doxastic nature. There are precedents in the literature of interpersonal trust that consider trust to be belief-based.<sup>14</sup> In the literature of trust in technology, and in particular AI, doxastic accounts are even more prominent.<sup>15</sup>

However, definition 3.1 presents certain problems. Tentatively, I have defined trust as a belief. This approach relegates trust to the mental realm, even though trust is strongly associated with action. Even though trust beliefs are the base that enables potential delegation, it does not seem entirely correct to put the weight of the definition on the trust belief. Doing so makes potential delegation appear as some kind of side effect. But, on the contrary, potential delegation is essential for trust. This nuance is subtle but relevant.

Trust beliefs are necessary in order to trust. But there seems to be something else. Often, people believe in propositions without those beliefs triggering any action. A trust belief does not imply delegation, but there exists a tight link with the possibility of it. Potential delegation is not an accident in trust. Instead, I argue that potential delegation is where the focus of the definition of trust should be. In order to address this issue, I will use the notion of disposition.

### 3.3.2 Trust as a belief-based disposition

#### Dispositions

A disposition is a state or condition that, under certain circumstances, allows further specific behaviour (Mumford 1998, pp. 20–22). This way, the term disposition is used to describe states which not only refer to the actual state of the world but to possible states that could follow from the current one. For example, an object being

<sup>14</sup>See Adler (1994), Hardin (2002), Hieronymi (2008), and Keren (2014).

<sup>15</sup>See Ferrario, Loi, and Viganò (2020), Castelfranchi and Falcone (2010), Jacovi et al. (2021), and Shin (2021).

soluble means that it would dissolve *if* it is submerged in water. Thus, solubility is a disposition. I argue that trust presents a similar structure: a trustor trusting a trustee means that the former would delegate something to the latter *if* the trustor wished or needed to do so. I proceed to develop this idea.

Dispositions can be characterized by having a *stimulus condition* and a *manifestation* (Choi and Fara 2012). In a disposition, the stimulus is the trigger for a specific behaviour. Such behaviour is the manifestation. In the case of solubility, the stimulus condition is putting an object  $x$  in water, and the manifestation is  $x$  being dissolved. Thus, dispositions refer not only to the current state of the world but to states of the world in which some condition obtains, such as  $x$  being put in water. That is why dispositions are commonly analysed in terms of counterfactual conditionals, which are statements of the sort “if  $X$  had been...  $Y$  would have been...” (Lewis 1973, pp. 59–60). In the middle of the 20th century, the so-called Simple Conditional Analysis (SCA) of dispositions became fairly popular in epistemology (Ryle 1949, pp. 31–33; Goodman 1954, pp. 87–90; Quine 1960, pp. 203–207). The analysis is as follows:<sup>16</sup>

**Definition 3.3.** *Disposition (SCA).* A disposition is a property  $D$ , of an object  $x$ , that is characterized by an associated stimulus condition  $C$  and manifestation  $M$ . The object  $x$  possesses  $D$  iff  $x$  would  $M$  if it were the case that  $C$ .

Despite its popularity, SCA has faced a variety of criticisms.<sup>17</sup> In what concerns this thesis, I would like to focus on one in particular. Dispositions have raised the interest of many philosophers in the past century.<sup>18</sup> Now the reader may wonder, what does all have to do with trust? As we have seen, trust does not only refer to the actual world but to hypothetical ones. Trusting a trustee means being open to delegating some task to them, under certain circumstances; namely, that the trustor wanted or wished to do so. This phenomenon matches the structural workings of a disposition: if the stimulus  $C$  of needing or wishing to delegate  $\delta$  were the case, then the manifestation  $M$  of delegation would occur. In other words, when the trustor trusts, they would delegate if they wished or needed to do so: a trustor  $x$  would  $M$  if they  $C$ -ed. This way, ‘trusting’ would operate similarly to dispositional properties as ‘being soluble’ or ‘being poisonous’.

This seems to match SCA, as presented in 3.3. However, treating actual delegation as the manifestation of trust does not come without problems. I proceed to show why. Let us hold the three following assumptions: 3.3, that trust is a disposition and that its manifestation ( $M$ ) is actual delegation. If  $M$  is actual delegation occurring, reliance could fit into 3.3 too. As we saw in chapter 1, reliance also results in delegation (while the subject wishes or needs to delegate).<sup>19</sup> Hence, it is possible to redefine reliance in SCA terms as follows: the subject  $x$  *relies on*  $B$  iff  $x$  would

<sup>16</sup>Definition 3.3 is an indirect quote from Choi and Fara (2012). For further details about the concept of disposition, their entry poses a great overview.

<sup>17</sup>The most notable critiques refer to finking, masking and mimicking situations. For details, see Martin (1994), Johnston (1992, pp. 233–34), Bird (1998), and Lewis (1997).

<sup>18</sup>In this thesis, I focus on the most recent discussion on dispositions. However, the concept has raised philosophical interest for much longer.

<sup>19</sup>Note that it is not my aim to characterise reliance as a disposition. I include the parenthesis about the wish/need of delegation to make the parallel with trust clearer, but in the case of reliance,

delegate  $\delta$  to  $B$  if  $x$  wished or needed to do so, meaning that  $x$  would  $M$  if they  $C$ -ed. Furthermore, because definition 3.3 uses a simple conditional, it allows the possibility of  $M$  occurring without  $C$  being the case; that is,  $x$  relying on  $B$  even without actually wanting to do so. This is different from trust (see section 1.6). But understanding actual delegation as a dispositional manifestation  $M$  blurs the line between trust and reliance, since both trust and reliance could fit into the definition 3.3 of disposition. I have already pointed out why I consider that lack of conceptual distinction undesirable.

In conclusion, dispositions seem to conceptually clarify some traits of trust. However, characterizing trust as a disposition makes unclear its distinction from reliance. Something seems to be missing in the trust characterization as a disposition. The key difference between trust and reliance has been outlined in section 3.2.3:<sup>20</sup> while reliance can occur with or without believing that the trustee is likely to succeed at the delegated task, trust requires something else. Trusting implies believing that the trustee is trustworthy.

### Redefining trust as a belief-based disposition

Rather than seeing the conceptual limitations of dispositions to capture trust in all its richness as a dead end, I consider them as an opportunity for refinement. I argue that trust is indeed a disposition; just not any kind of disposition. While sharing structural similarities with concepts such as solubility or toxicity, the concept of trust seems to go beyond counterfactual delegation. I will argue that the missing element is trust belief, which leads me to redefine trust as a belief-based disposition:

**Definition 3.4.** *Trust* is the belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold

- (i)  $A$  has a continuing belief that  $B$  is trustworthy.
- (ii) If  $A$  wished or needed  $B$  to do  $\delta$ , then  $A$  would be willing to delegate  $\delta$  to  $B$ .

Typically, a disposition  $D$  is defined following the structure “ $x$  is  $D$ -able if  $x$  is disposed to  $M$  when  $C$ ”. In the definition above, I use a different structure to highlight that trust requires an additional condition besides the conditional relationship between  $M$  and  $C$  (condition (ii)): trust belief (condition (i)). Nevertheless, if one really wanted to express trust in terms of SCA (see 3.3), that is possible. Such definition would read as “ $A$  trusts  $B$  iff  $A$  is disposed to delegate  $\delta$  to  $B$  when  $A$  wishes or needs to do so (as long as  $A$  holds a trust belief about  $B$ )”. For clarity, I will stick to the formulation 3.4.

There are different elements involved in the definition above. I proceed to unfold them bit by bit. Let us start with condition (i): continuing trust belief. Condition (i) serves two main functions. The first one is to highlight the role of trust belief as a necessary condition for trust. In other words: trust is belief-based. By belief-based,

---

such a wish/need is not a condition that refers to hypothetical scenarios. The subject wishes or needs to delegate (and in fact does delegate) necessarily in the actual world. If not, there is no reliance at all. This necessary actuality does not fit the counterfactual nature of disposition.

<sup>20</sup>Further details in section 1.6 of chapter 1.

I mean that trust is a phenomenon that cannot occur without the trustor holding a trust belief about the trustee. The reason is that trust is a concept that does not only refer to the material reality of how things are and could be; it also possesses a cognitive charge. Several authors<sup>21</sup> back up the idea that trust is a cognitive concept. They hold what in the literature is often called a doxastic account of trust. I agree with doxastic accounts of trust, that trust entails trust belief. Reliance, on the other hand, does not. It is possible, and in fact common, to delegate tasks to others without much thought. Sometimes, we delegate tasks to people who we do not consider trustworthy. If we are lucky, those people may perform the task under our command successfully. However, without trust belief, that is not trust but reliance. Condition (i) enables me to make such a distinction. The second purpose of (i) is to ensure that  $A$  holds the trust belief in every possible world used to evaluate (ii). That is, the trust belief *continues* over all the relevant hypothetical scenarios, namely, every scenario in which  $A$  wishes or wants  $B$  to do  $\delta$ . Opposed to continuing beliefs there are discontinuing beliefs; that is, beliefs that the belief holder ceases to have at some point in time. For example, imagine that I used to believe that my neighbour was trustworthy until I caught him in a lie. The moment I discovered he was lying, I stopped believing he was trustworthy. Hence, my trust belief is not continuing (thus condition (i) is not met, therefore I do not trust my neighbour). This way, trust belief operates as a background condition for trust. Condition (i) is almost equivalent to definition 3.1. I criticized that definition due to its lack of connection between trust beliefs with the actual act of delegation. Something seemed to be missing. I proposed the concept of disposition to fill that gap. That takes us to condition (ii).

Condition (ii) refers to the dispositional character of trust. Trusting a trustee does not mean just delegating something to the trustee, but believing that delegation is a reasonable option in some hypothetical future scenario (hence the use of the progressive conditional in ‘would be willing’). When the trustor trusts, they are opening themselves to potentially delegate some task to the trustee. They do so because they believe that the trustee is trustworthy; that is, they have reasons to believe that the trustee will perform the task successfully, moved by motives that the trustor deems appropriate. This is what I call a *belief-based disposition* (or BBD for short).

**Definition 3.5.** *Belief-based disposition* refers to the kind of dispositions that originate in belief. That means that belief-based dispositions are states in which an object  $x$  is disposed to  $M$  when  $C$  only under the background condition of a certain belief being held.

The concept of belief-based disposition is quite narrow because it is inspired specifically by the conceptual workings of trust. Since they require belief, belief-based dispositions refer to certain kinds of relationships between agents among whom, at least one of them, needs to hold a certain belief. In the case of trust, it would be a trust belief. However, trust is not the only example of this phenomenon. Take manipulation as an alternate example. A subject  $B'$  being manipulated means that

<sup>21</sup>See Adler (1994, pp. 270–72), Hardin (1993, p. 152), Hieronymi (2008, pp. 221–24), Keren (2014, pp. 31–35), Ferrario, Loi, and Viganò (2020, p. 532), Castelfranchi and Falcone (2010, p. 35), Jacovi et al. (2021, p. 625), and Shin (2021, pp. 24–25).

they would act according to the interests of a different subject  $A'$ , if  $A'$  asked  $B'$  to do so. Thus, we see how manipulation has a dispositional character. However, such a character is not enough to define the concept. It is not hard to think about possible situations in which  $B'$  would please  $A'$  without being manipulated.  $B'$  could simply be altruistic, for example. Because of this, belief appears as a necessary background condition, in order to distinguish manipulation from other concepts in which its stimulus and manifestation also take place. As with trust, holding a certain belief is crucial. In this case, the ‘manipulation belief’ would be some belief induced in  $B'$  by  $A'$  to  $A'$ ’s own benefit (which makes  $B'$  act as they do). Thus, we can conclude that manipulation is another example of a belief-based disposition.

There is still one piece missing for definition 3.4 to be complete. I have insisted on the importance of condition (i) and how a trustor cannot trust a trustee without believing that the trustee is trustworthy. The natural follow-up question is: what does it mean to be trustworthy? To answer this, definition 3.2 suffices:  $B$  is trustworthy (for  $A$ ) if it is reasonable for  $A$  to delegate  $\delta$  to  $B$ . Note that the difference between being trusted and being trustworthy is subtle but important. Being trustworthy implies the existence of rational reasons that allow  $A$  to conclude that they could safely delegate  $\delta$  to  $B$ . In contrast, being trusted does not require such reasons. Being trusted by  $A$  only means that  $A$  *believes* that they can safely delegate, but  $A$ ’s belief may be grounded on rational reasons or not.

### Appropriate trust

I have redefined the concept of trust according to my BBD account. The aim of this new definition is to capture the conceptual workings of trust, independently of the trustee’s nature (personal, institutional or artificial). Thus, definition 3.4 describes how trust works, without going into details on how it *should* work. I propose the term ‘appropriate trust’ to refer to the kind of trust that is normatively desirable.

**Definition 3.6.** *Appropriate trust* is the justified belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold

- (i’)  $A$  has a continuing *justified* belief that  $B$  is trustworthy.
- (ii’) If  $A$  wished or needed  $B$  to do  $\delta$ , then  $A$  would be willing to delegate  $\delta$  to  $B$ .

The difference between definition 3.4 and 3.6 is that the latter requires *justified* trust belief. What I aim to highlight with (i’) is that holding a trust belief about the trustee is not enough to trust them appropriately. In the case of appropriate trust, the trust belief must be *justified*. That is, the trust belief should be supported by previous beliefs that are justified in turn. Here, I am relying on a coherentist view of justification, according to which a belief  $b$  held by an agent  $X$  is justified iff such a belief fits  $X$ ’s belief system.<sup>22</sup> I use ‘fit’ in a wide sense, meaning that  $b$  does not enter in contradiction with the rest of  $X$ ’s beliefs, but also that the epistemic attitude of  $X$  towards the content of  $b$  (in this case, their degree of belief) is consistent with the

<sup>22</sup>For a detailed view on a coherentist notion of epistemic justification see Lehrer [1990] p. 89; Elgin [2005] p. 158; Olsson [2021]; Levi [1980] pp. 1–2.

rest of their epistemic attitudes.<sup>23</sup> This way, a coherent set of beliefs is conformed by beliefs that are supported by each other, meaning that a belief  $b_1$  is supported by a belief  $b_2$  iff  $b_1$  is more credible if  $b_2$  is true than if it was false (Lewis 1946, p. 338; Spohn 2018, p. 3840).<sup>24</sup>

A coherentist view of justification suits the picture of trust I aim to paint in this thesis. The reason is that trust is a complex phenomenon which finds its roots in a variety of places. People trust different kinds of trustees for different kinds of reasons. Thus, the quest to identify the source of justification for trust beliefs is a challenging endeavour –as it is to argue where lies belief justification in general. Therefore, when defining appropriate trust, I set more attainable goals: to prevent the absence of clear reasons to trust and to avoid contradiction, both internal and, ideally, with external sources. Let me illustrate how this notion of justified trust belief works with an example.

Once again, picture Claire and Christian from case study C (figure 3.1).<sup>25</sup> How could be Claire be justified in trusting Christian? Let us say that Claire holds the following beliefs:

- ( $b_1$ ) Claire believes that Christian has several cats.
- ( $b_2$ ) Claire believes that Christian treats his cats very well.
- ( $b_3$ ) Because of ( $b_1$ ) and ( $b_2$ ), Claire believes that Christian loves cats.
- ( $b_4$ ) Claire believes that ( $b_3$ ) acts as a motive  $M$  for Christian to take good care of her cat if she asked him.
- ( $b_5$ ) Therefore, Claire believes that Christian is trustworthy: it is reasonable for her to delegate to Christian to take care of her cat.
- ( $b_6$ ) Furthermore, believing ( $b_3$ ) makes Claire believe that Christian is a good person.

In the above example, ( $b_5$ ) is the trust belief, which it is justified by its coherence with beliefs ( $b_1$ )-( $b_4$ ). Everything that Claire believes about Christian is compatible with her conclusion that Christian is trustworthy. Therefore, Claire's trust belief is justified and her trust in Christian is appropriate. Trust beliefs are part of the trustor's belief system: they are supported by other beliefs (( $b_1$ )-( $b_4$ )) and they can trigger additional beliefs in turn (( $b_6$ )). That means that, to be justified, the trust belief needs to cohere with the web of belief that the trustor holds. That is, it fits with other beliefs in a way that makes the entire system of beliefs internally consistent and mutually supportive. This way, internal contradictions are avoided. Regarding external data –such as the fact that Christian has several cats–, let us assume Lewis'

<sup>23</sup>I will stick to the general terminology used in Gärdenfors (1988); here, in particular, I am following the notion of epistemic attitudes presented in pages 12-13.

<sup>24</sup>Here, I have attempted to apply Spohn's notion of *credibility* to Lewis' *congruence*, which is equivalent to what I intuitively understand by coherence. Note that in his paper, Spohn (2018) talks about  $b_1$  being a reason for  $b_2$ , rather than  $b_1$  justifying  $b_2$ . The nuances and differences between these kinds of epistemic relationships go beyond the scope of the present thesis.

<sup>25</sup>See section 2.4.3 in chapter 2.

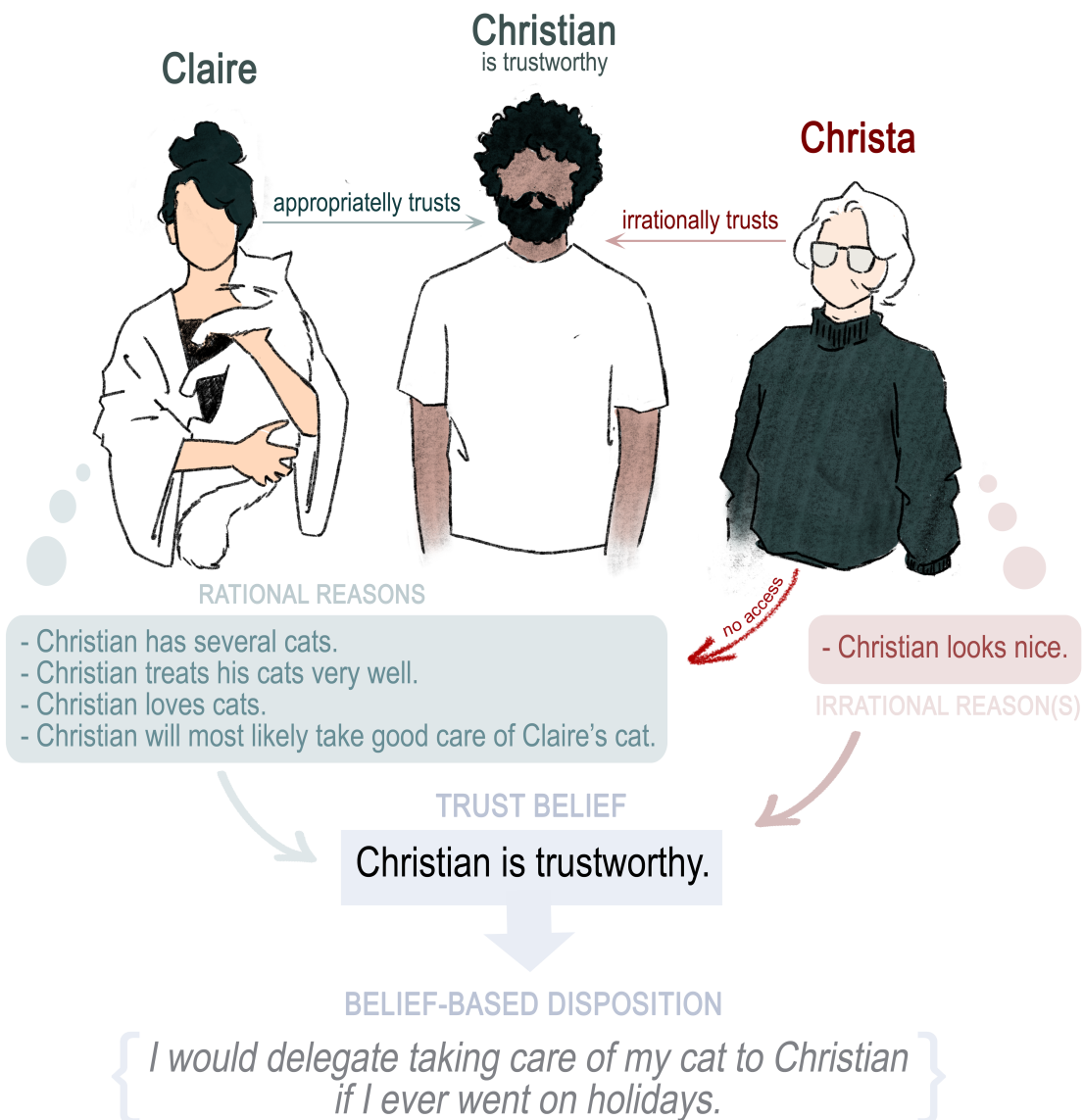


Figure 3.1: In this example, different trustors (Claire and Christa) trust the same trustee (Christian). However, trust can be the fruit of different reasonings which can be rational, or not. When the trustor trusts due to rational reasons, as Claire does, their trust is appropriate. When the trustor's reasons to trust are irrational, as in Christa's case, trust is unwarranted. Trust can be unwarranted even when placed on trustworthy trustees; what makes trust (un)warranted is whether the trustor has access to rational reasons that allow them to conclude that they could safely delegate a certain role or task to the trustee.

Principal Principle: a rational trustor conforms their degrees of belief to objective chances (Lewis 1994, p. 475). That is, under normal circumstances,<sup>26</sup> and given that Christian’s testimony does not enter in conflict with previous Claire’s beliefs, it is rational of Claire to believe Christian’s testimony about his cats. In other words, if Christian told Claire that he has cats, and she lacks further reason to doubt this is true, it is rational to believe that it is probably true. Thus, having established this connection between the trustor’s degrees of belief and the world, a justified trust belief is coherent not only with the rest of the trustor’s beliefs but also with external sources, whilst accessible to the trustor.

The notion of appropriate trust is an important element of the proposed BBD account on trust, and it becomes key when the trustee is an AI system. The reason is that subjective reasons should not be enough to trust AI systems, especially when used in high-stakes situations. As I pointed at the end of section 3.2, trust in AI demands the possibility of rational evaluation of the reasons to trust. Thus, trust in AI should be not purely subjective but *intersubjective*. That is the case for appropriate trust since justified trust belief allows different trustors to compare their trust belief and the beliefs that support them in turn. Let me illustrate this with a different example, this time having an AI system as a trustee.

Right now, I find myself writing this thesis. I am not a native English speaker, so I am unsure whether the phrase “the quest *to* identify the source of justification” is correct. I am doubting between the prepositions ‘to’, ‘for’, or maybe an entirely different formulation. So I ask ChatGPT (see figure 3.2). ChatGPT confirms that my phrasing is correct and, since I trust it, I end up including it in the thesis. I would say that my trust belief is justified and, therefore, my trust is appropriate. But what does that mean? That means that I hold the following beliefs:

- ( $b_1$ )’ I believe that “the quest *to* identify the source of justification” could be a correct formulation in English.
- ( $b_2$ )’ I believe that my English is not perfect and that I make mistakes sometimes.
- ( $b_3$ )’ I believe that ChatGPT is able to recognize patterns from a vast amount of text data and use those patterns to predict and generate text that is grammatically correct and contextually relevant.
- ( $b_4$ )’ Because of ( $b_3$ )’, I believe that ChatGPT is trustworthy: it is reasonable for me to delegate text generation tasks to it.
- ( $b_5$ )’ Furthermore, believing ( $b_2$ )’ and ( $b_4$ )’ makes me believe that ChatGPT’s sentences are more likely to be grammatically correct than my own.

I am justified in believing ( $b_4$ )’, my trust belief, because it coheres with beliefs ( $b_1$ )’, ( $b_2$ )’, ( $b_3$ )’ and ( $b_5$ )’. I do have access to external sources from where I have learnt how ChatGPT works, and I believe that such workings are likely to produce correct English sentences. Therefore, I trust the system to correct my grammar. Thus, ( $b_4$ )’

<sup>26</sup>An example of an abnormal circumstance would be the fact that Claire is schizophrenic and Christian has been an hallucination all along, rather than a real neighbour.

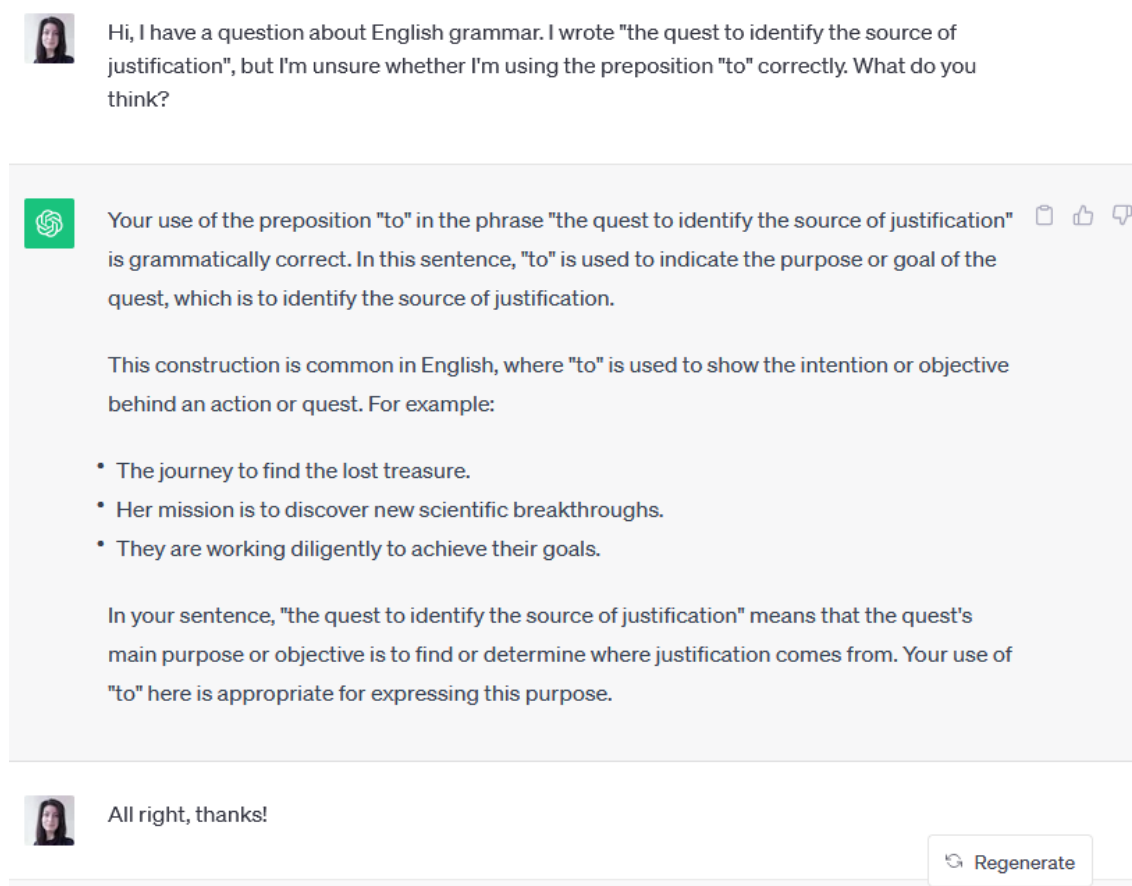


Figure 3.2: In this example, I trusted OpenAI's ChatGPT-3 to correct my grammar. (<https://chat.openai.com/chat>).

goes beyond my own subjectivity since it is supported by beliefs that can be shared and contrasted by others. If I got to receive new information, such as that Open AI used corrupted data that often leads ChatGPT to produce incorrect sentences,  $(b_4)$ 's coherence with my belief system would decrease,<sup>27</sup> decreasing my degree of belief in  $(b_4)$  in turn, and, in consequence, decreasing my trust in ChatGPT.

The example above illustrates the importance of establishing appropriate trust in the context of AI. When trust belief is justified, it is coherent with the trustor's belief system and, indirectly, with external sources, with real world data. This connection with the world enables the comparison of trust belief across various trustors, rendering appropriate trust intersubjective. Intersubjectivity becomes key when the trustee is an AI since intersubjectivity makes it more likely to place trust in trustworthy trustees. The reason is that, because trustworthiness refers to hypothetical behaviour, making it challenging to ascertain a trustee's trustworthiness definitively. However, having multiple sources to predict such behaviour enhances the accuracy of trustworthiness assessments. Going back to the above example: if my belief  $(b_4)$  is shared by different trustors and all of them have different belief systems with

<sup>27</sup>Because my belief system has been updated to incorporate a new belief,  $(b_6)$ , that GPT's training data is corrupted and to, therefore, revise beliefs  $(b_3)$ – $(b_5)$ .

which  $(b_4)$  coheres, the likelihood of  $(b_4)$  being true increases. Having trust belief validated by diverse trustors enhances its coherence and therefore its justification. While justification does not equate to truth, the higher the degree of justification the greater the chances of the content of the trust belief being true.

In sum, appropriate trust refers to the phenomenon of a trustor who is justified in believing that it is reasonable to delegate a certain task or role to a trustee. In other words, the belief that the trustee is trustworthy is coherent with the trustor's belief system, including their beliefs about external data that could be contrasted by different trustors. Appropriate trust is normatively desirable because it poses more chances to place trust in trustworthy trustees. This is the goal for any kind of trustee, but it becomes particularly relevant when it comes to artificial trustees. When we trust artificially intelligent systems, we expect them to achieve successful results, using methods that we deem appropriate (such as pattern recognition to generate text). Such an expectation is more likely to be fulfilled if it is supported by a coherent set of beliefs that are not purely subjective.

### 3.3.3 Challenges check

In section 3.2.1, I highlighted some features of trust that pose a challenge for most accounts, since it is not easy to capture such features simultaneously. Let me briefly revisit that list<sup>28</sup>

- (4) Trust can (and should) be motives-based (which makes it distinct from reliance).
- (5) Trust can (and should) be rational.
- (6) Trust can (and should) be intersubjective.

The question now is, does characterizing trust as a belief-based disposition resolve these challenges? I claim that the answer is positive.

Point (4) is resolved by addressing the trustee's motivation to act in the definition of trustworthiness (3.2). As I said, a trustee is trustworthy if it is reasonable to believe not only that they will succeed at  $\delta$ , but also that they will do so moved by motives deemed appropriate by the trustor.

The same applies to point (5). Trust is not always rational. But it is when placed on trustworthy trustees which, ideally, should be the case. When it is, trust is appropriate.

Point (6) deserves a little more attention. In the last section, I have tackled intersubjective (dis)agreement via examples, explaining how appropriate trust opens the door for intersubjectivity. Let me address the topic more systematically. Let us say that we have two different trustors,  $A$  and  $A'$ , and a trustee  $B$ .  $A$ 's personal experience with  $B$  leads them to believe that  $B$  is trustworthy. However, there is a piece of information that enters in conflict with  $B$ 's presumed trustworthiness.  $A'$  has access to this extra information, while  $A$  does not. But since both  $A$ 's and  $A'$ 's belief systems are each coherent,  $A$  and  $A'$  can compare such systems and the specific beliefs that supported the belief that  $B$  is un/trustworthy. Thus,  $A$  can access the

---

<sup>28</sup>Once again, the list begins at (4) for consistency with chapter 2 and throughout the present chapter.

extra information through  $A'$  and re-evaluate their own trust belief. This way, it is possible that both trustors end up agreeing that  $B$  is untrustworthy.

Note that the case presented above is an idealization. It is possible that trustors  $A$  and  $A'$  have originally identical belief systems but that their respective belief revision processes lead them to different conclusions.<sup>29</sup> Nevertheless, even in that case, intersubjectivity remains highly valuable. Even if  $A$  and  $A'$  contrast their trust belief, share their sources, compare their belief systems and still agree to disagree, the possibility of holding such a discussion already enriches the epistemic background of their respective trusting relationships. That is, given that  $A$  maintains their trust belief, such a belief is better supported now that  $A$  has had access not only to their own belief system's but to  $A'$ 's. In turn,  $A'$ 's distrust towards  $B$ , now that not even  $A$ 's reasoning could convince  $A'$  that  $B$  is trustworthy, can only grow stronger. Therefore, even if different trustors' exchange ends up in epistemic disagreement, the possibility of disclosing trust beliefs intersubjectively remains beneficial.

In sum, my aim with the paragraphs above was to convey the importance of appropriate trust in the AI field. For trust (in artificial trustees) to be appropriate, trust must be motives-based, grounded in justified trust belief and intersubjective. For these criteria to be met, both technical and social aspects are relevant, which matches the socio-technical nature of AI. For example, a necessary condition for trust belief to be justified when the trustee is an AI is that the system has to be technically capable of performing the task that is being trusted with. In addition, it is important to consider how this technical capability is communicated to different trustors, so they have the chance to reflect on it and have an intersubjective discussion about what the system should (and should not) be used for. AI is a socio-technical tool, which means that “the trustor” does not only refer to individual users but to society as a collective. Being able to discuss in a social frame the reasons to trust or distrust AI is normatively desirable. By taking a variety of angles into account, the acceptance of AI in society would be better contrasted, more democratic and a further scrutinized social position. Since AI affects us all, these demands are reasonable and fair.

---

### 3.4 Possible objections

I have devoted this chapter to developing my proposed BBD account of trust. In this section, I will address parts of the literature on trust that enter into direct conflict with my views.

The BBD account can be considered an impure doxastic account of trust. I take the label from Keren (2020, p. 109), who defines impure doxastic accounts of trust as those accounts according to which trust belief is necessary but not sufficient for trust to materialise. Some additional condition is needed. The additional condition that I propose in the BBD account is (ii).

---

<sup>29</sup>The above can happen due to  $A$  and  $A'$  using different plausibility measures. For details, see Spohn (1988, pp. 112–116; Halpern (2003, pp. 43–45

In contrast to doxastic accounts, we can find non-doxastic ones. In those accounts, trust beliefs are not considered necessary to establish a trusting relationship between the trustor and the trustee. Authors such as Jones (1996) and McLeod (2021) defend that trust is an affective attitude. Other non-doxastic accounts suggest that trusting involves adopting a moral stance towards the trustee, being the beliefs about them irrelevant (Holton 1994, p. 63). Alternatively, there are authors who characterise trust as a disposition to rely on a trustee, without necessarily holding any kind of mental state (Kappel 2014, p. 2017). All these positions bring valuable aspects to the debate on trust. However, insofar as they are non-doxastic, these contributions enter into direct conflict with my account of trust as a belief-based disposition. I introduced trust belief as a necessary condition for trust because, without it, trust would be practically equivalent to reliance. However, we have seen in chapter 1 that the difference between the two concepts is significant. If trust belief is neglected, trust is defined as a relationship based on emotions, affections or morality alone. However, this leaves out many relevant cases of trust, such as institutional trust or trust in AI. Therefore, I argued that doxastic theories of trust are more complete than non-doxastic ones and they capture the phenomenon of trust more accurately.<sup>30</sup>

Defendants of non-doxastic views on trust could argue that my account, as any other doxastic one, does not explain therapeutic trust (Keren 2014, p. 8). Therapeutic trust refers to situations in which the trustor does not believe that the trustee is trustworthy, but decides to trust them anyway with the aim to encourage some behaviour in the trustee. A typical example is the parents who trust their not-so-responsible daughter to stay home alone, so she gets used to assuming a certain degree of responsibility. Keren (2020, pp. 116–7) offers a reply to this argument, pointing out the fact that trust, belief and trustworthiness come in degrees.<sup>31</sup> In the case of therapeutic trust, the trustor believes, to some degree, that the trustee has a degree of trustworthiness. No parents would leave home alone their daughter if the daughter was a newborn baby, mentally impaired or consistently and purposely irresponsible; in other words, in case they did not believe *at all* that she is trustworthy. The idea behind therapeutic trust is that the trustor acting *as if* they trusted will promote the trustworthiness of the trustee; that is, will increase their degree of trustworthiness. Note the emphasis in ‘as if’. As I conceive trust, the so-called therapeutic trust is not really trust, but some sort of simulation of it. I do not consider this a threat to the BBD account because I do not consider therapeutic trust as trust, but rather as some kind of reliance, which indeed can be the case without implying trust belief.

There are also authors who have proposed accounts of trust with artificial trustees in mind, reaching different conclusions than the one I provided here. Ferrario, Loi and Viganò’s (2020) work is particularly interesting. They propose a theory that aims to capture both trust in AI and interpersonal trust. Their model consists of three components: simple trust, reflective trust and paradigmatic trust. According to them, simple trust equals reliance since it refers to the intention of a trustor to delegate without implying trust belief. At first glance, reflective trust seems similar

---

<sup>30</sup>See section 3.2.3.

<sup>31</sup>For details, see section 1.5 in chapter 1.

to my general definition of trust, and paradigmatic trust might appear equivalent to what I have called appropriate trust. According to them, trust is the trustor’s belief that the trustee is trustworthy. They define trustworthiness as follows:

“*Y* is trustworthy for *X* to do *A* for the sake of *G* =<sub>def</sub> *Y* has properties that are objective reasons for *X* to simply trust *Y* to perform *A* to reach the goal *G*. This is equivalent to state that *Y* has properties that are objective reasons for *X* to rely on *Y* in the absence of control” (pp. 531-2).

There are two main differences between my approach and Ferrairo’s. The first one lies in the definition of trustworthiness, as the quote above illustrates. While Ferrairo, Loi and Viganò talk about objective reasons, I postulated the existence of rational reasons and emphasized coherence instead.<sup>32</sup> Their quote seems to imply that it is possible to externally evaluate whether a trustor *A* should trust a trustee *B*. However, I do not think that such an evaluation is possible. I consider the notion of ‘objective reasons’ to be problematic due to the subjective nature of trust. In contrast, I described trustworthy trustees as those about whom it is reasonable to believe that they are likely to successfully perform a certain task. I used the phrasing ‘reasonable for the trustor’ with the aim to refer to the subjective reasoning that makes a trustor trust. I argued that this reasoning cannot be objectively evaluated. It can be discussed by others and intersubjectively (dis)agreed upon, but, ultimately, there is no way of determining who should trust whom.

A second key difference between Ferrairo’s account and my own is that they propose a model based on delegation,<sup>33</sup> rather than to the possibility of it. I have emphasized why trust is based on the latter rather than on the former. In case actual delegation never materialises, it is still possible to hold a trusting relationship towards the trustee. A doxastic view on trust like the one proposed by Ferrario, Loi, and Viganò (2020) seems to ignore the dispositional aspect of trust. In order to avoid that, I have defined trust as a belief-based disposition, rather than as belief-based reliance. I consider this to be a crucial distinction. Despite my disagreement, I consider Ferrario, Loi and Viganò’s (2020) model very relevant since it puts the focus on the fact that trust comes in degrees and poses a formal explanation of how that happens. In addition, it does a good job of bringing together different kinds of trust placed in different kinds of trustees. In the case of trust in AI, this is a very valuable contribution to the field.

Finally, I would like to allude to another publication which shares the purposes of this thesis, but proposes a different approach. Jacovi et al. (2021) also propose a theory of trust with the aim to capture the phenomenon in human-AI interaction. Their focus is on anticipation and vulnerability. In particular, they define trust in AI as the human trustor’s belief that a set of contracts will be upheld. Such contracts refer to “any functionality that is deemed useful, even if it is not concrete performance at the end task that the model was trained for” (p. 626). Like myself, they also propose a view on trust that characterises it as belief-based. But unlike the BBD account, their description of trust does not address what trustors would be willing

---

<sup>32</sup>See definition 3.2

<sup>33</sup>In their words, “perform *A* to reach the goal *G*”, as already quoted.

to do when involved in a trusting relationship. Their proposal can be classified as doxastic and, on this basis, my critique from section 3.3.1 holds: by identifying trust with trust belief entirely, the possibility of delegation is relegated to a mere accident. As argued previously, I disagree with that idea. I bring Jacovi et al.'s (2021) work to the reader's attention because, despite presenting the same problem as other doxastic theories of trust, it poses a good representation of the literature in which features of interpersonal trust have been adapted to an account of trust in AI specifically.<sup>34</sup>

In addition, Jacovi introduces the idea that trust in AI can be fostered through explanations. If the user receives an explanation of how the system works, and as a consequence, the user believes that the system will fulfil their expectations, then they trust the AI. This is an intuitive and widely agreed claim in the literature of explainable AI.<sup>35</sup> In Jacovi's work, the claim works for their account on trust. In the next chapter, I will investigate whether explanations do indeed foster trust in AI in the frame of the BBD account.

---

<sup>34</sup>Other attempts can be found in Lin, Wang, and Hajli (2019) and Yan et al. (2013).

<sup>35</sup>It is an assumption present in Creel (2020), Ribeiro, Singh, and Guestrin (2016), Miller (2017), Goebel et al. (2018), Hoffman, Klein, and Mueller (2018), Gilpin et al. (2018), Anjomshoae et al. (2019), Wang et al. (2019), and Shin (2021), among others.



# 4

---

## *Building trust*

---



---

## 4.1 Introduction

Trust, like most things of value, is hard to build but fairly easy to break. While it takes time and perseverance to prove oneself trustworthy, a single failed commitment can permanently tarnish one’s reputation. That is the paradox of trust, which, as I aim to illustrate, extends to the realm of AI.

In previous chapters, I have explored the nature of trust, drawing a parallel between trust in humans and trust in AI. In chapter 3, I proposed an account of trust as a belief-based disposition that encapsulates the potential for diverse trustees, both human and artificial. In a nutshell, in the BBD account, trust is the disposition of the trustor to delegate some task or role to the trustee. Such a disposition originates in the trustor’s trust belief, which is necessary for a trusting relationship to be developed. With this account, I have shed some light on what trust is. In this chapter, I aim to answer the question of how trust can be built, in particular, in the case of artificial trustees.

When it comes to people, there are a variety of reasons for which building trust can be valuable: to enhance cooperation, to manifest respect, to develop meaningful attachments, etc. In the case of artificial trustees, the value of trust appears indisputable across ethics guidelines and white papers.<sup>1</sup> In particular, the European Commission states that “trustworthy AI can improve individual flourishing and collective wellbeing by generating prosperity, value creation and wealth maximization” (HLEG 2019, p. 9). These kinds of phrasings highlight how promising trustworthy AI appears to be (to some), without being too specific about which purposes trust should serve and, accordingly, how trust should be built.

It is however not obvious how to design trustworthy AI systems. One of the biggest challenges when researching whether AI should be worthy of trust is the black box problem. The black box problem refers to the fact that many AI systems –including NNs<sup>2</sup>– are opaque, meaning that it is not possible for human beings to understand how the system reaches its outputs. The term ‘black box’ can refer to a variety of systems since there exists different kinds of opacities. While technical opacity –where the inner workings of a system are not accessible or understandable– has been widely discussed in philosophy (Creel 2020; Zerilli et al. 2019), it is not the only form of opacity that can affect trust. As Vredenburg (2022, pp. 83–84) notes, some systems are deliberately opaque to protect company secrets. Additionally, opacity can arise from technical illiteracy, where certain systems are inaccessible to individuals without the necessary technical background (Burrell 2016, p. 4). Regardless of whether the opacity stems from inaccessible algorithms, implementation in code, the training dataset, or a lack of technical expertise, this inaccessibility generates opacity, which poses significant barriers to cultivating trust. In recent years, explainable AI (XAI) has appeared as a promising option to foster trust in AI systems.

The literature on AI<sup>3</sup> often assumes that explanations lead to trust. This assumption has been referred to as the *Explainability-Trust Hypothesis* (ET) (Kästner

---

<sup>1</sup>OECD (2017), HLEG (2019), Leslie (2019), United Nations (2021), and Gunning et al. (2019).

<sup>2</sup>For details, see section 2.2

<sup>3</sup>Creel (2020) and Ribeiro, Singh, and Guestrin (2016).

et al. (2021). It is common to use ET to argue for XAI. However, the relationship between explanations and trust is complex and, in this chapter, I will argue that taking ET for granted is problematic. The existing literature on ET (or that takes ET as a background assumption) has overlooked several issues that demand thorough investigation. In the first place, it is unclear what are the additional conditions under which *some* explanations lead to trust. Secondly, it remains ambiguous what is meant by trust in this context. I aim to fill these gaps by shedding light on the conditions under which ET holds, building up on the account of trust offered in chapter 3. To do so, I will analyse ET's epistemic limitations (Blanco 2022). Such analysis underscores ET's context dependency. Specifically, this is the case regarding the target and quality of the explanations. Both the explanation recipient's<sup>4</sup> background knowledge and the quality of the offered explanations need to be taken into account to determine whether such explanations lead to trust. Since these nuances are not made explicit in ET, I will question the common assumption that explanations inevitably lead to trust, arguing that such a link is not guaranteed. A question left open is whether ET's limitations could be overcome. In this chapter, I offer a proposal to do so by reformulating ET in such a way that the nuances above are taken into account (ET\*).

In essence, ET\* expresses that explanations only lead to trust under certain circumstances. In other words, only *some* explanations lead to *some* degree of trust. Using the BBD (belief-based dispositional) account of trust and introducing a gradual notion of trust, I will argue that the gradual nature of trust can be explained if trust is understood as belief-based. In my view, different degrees of trust correspond to different degrees of trust belief. Thus, whilst –under some circumstances– explanations can influence trust belief, trust can be fostered by explanations. In addition, explanations can also foster distrust, which has a complementary value to trust when distinguishing trustworthy trustees from untrustworthy ones. In this chapter, I will provide a detailed analysis of how this works in the case of artificial trustees and, in particular, in the case of XAI.

The chapter is structured as follows. In section 4.2, I will start by discussing the value of trust (in AI) and why building it is important. In section 4.3, I introduce ET and to explore its epistemic limitations. As a result, I will propose a more nuanced new version, ET\*, with the aim of overcoming its predecessor's problems. In section 4.4, I will build on ET\* and the BBD account of trust to explain how degrees of trust correspond to degrees of trust belief. This variability in explanations becomes particularly advantageous for artificial trustees. This is because various AI applications require different degrees of trust from different trustors to be successfully implemented. Finally, in section 4.5, I will discuss the value of distrust as complementary to trust, putting the focus on artificial trustees.

---

<sup>4</sup>Term taken from Bordt et al. (2022).

---

## 4.2 The value of trust

As mentioned in previous chapters,<sup>5</sup> numerous ethics guidelines and white papers highlight the importance of trust in shaping human-AI relationships.<sup>6</sup> However, these documents often lack specificity when explaining why trust in AI is so valuable and what its purpose is exactly. In the context of human relationships, trust is valuable because it is considered a sign of respect, enhances cooperation, and fosters meaningful attachments (McLeod 2021; Hardin 2002; Kirton 2020). Some authors also argue that trust has intrinsic value in itself (O’Neil 2012, p. 311; McLeod 2021). In contrast, in the case of artificial trustees, however, the value of trust often seems to be viewed as instrumental. For example, the European Commission claims that “trustworthy AI can improve individual flourishing and collective well-being by generating prosperity, value creation, and wealth maximization” (HLEG 2019, p. 9).

In general, trust enhances cooperation, even when it is not deemed necessary for it (Cook, Hardin, and Levi 2005, pp. 166–170). In interpersonal relationships, trust removes or minimizes the need for the trustor to monitor the trustee, thereby facilitating smoother cooperation between the two. Since AI systems do not have interests of their own, I won’t go as far as saying that human users *cooperate* with them to achieve shared goals. Nevertheless, trusting the system that is being used, helps human agents to make the most of AI. Moreover, developing systems that surpass human abilities is of limited value if humans do not trust these systems enough to delegate tasks to them. In fact, when the trustee is an AI system, a trusting relationship facilitates engaging with such a system more openly and confidently. Trust in AI encourages users to rely on its functionalities, interact with it more comfortably, and embrace its recommendations or decisions with greater ease. Consequently, trust in AI becomes an indispensable requirement for its successful implementation in society (HLEG 2019, p. 14; Leslie 2019, pp. 5–6).

It is important to notice that trust is only beneficial if it is warranted. In the case of AI, this means that the goal is not mere trust, but the kind of appropriate trust that was introduced in chapter 3. Then, to pursue human-AI trusting relationships, the goal should be that the human trustor holds a justified trust belief. Following the argument presented in section 3.3.2, the goal of building trust in AI is for the trustor to believe that the system will most likely achieve successful results and that such a trust belief is coherent with the rest of their belief system. For such coherence to exist, the trustor needs to hold additional beliefs about system that align with their trust belief. This means the trustor needs access to additional information about the system, such as how it works, how it was trained or what its success rate performance is, for example. Accessing this kind of information can be challenging in some cases, especially when the system is a black box (Buijsman 2022, pp. 563–64). XAI is one attempt to solve the so-called black-box problem of AI.

In AI, opacity seems to be undesirable since it makes it harder not only to interpret the system’s results but also to resolve potential problems and detect bias.

---

<sup>5</sup>See introductions to chapters 1 and 2.

<sup>6</sup>OECD (2017), HLEG (2019), Leslie (2019), United Nations (2021), and Gunning et al. (2019).

As discussed above, it has been argued that trusting human-AI relationships favours an optimal use of AI systems, and with it, their successful implementation in society. It is often assumed that making AI systems explainable contributes to building trust in those systems (HLEG 2019, p. 13). This assumption only makes sense in the case that explanations lead to trust. I will refer to this premise as the *Explainability-Trust Hypothesis* (ET).

---

## 4.3 The Explainability-Trust Hypothesis

### 4.3.1 Background and formulation

Explainability refers to the property of a system whose outputs can be explained. In other words, explainable systems are thought to provide explanations that help humans understand how the system operates. It is often considered that explainability is a way to reach transparency, which is opposed to opacity. In this sense, a system is opaque if it reaches outputs using processes that cannot be understood by humans. Many of the AI systems that are currently being developed are opaque. This is the so-called black-box problem. XAI is thought to be a promising attempt to solve it.

XAI systems aim to offer tools that generate explanations for otherwise opaque AI systems (Buchholz 2023, p. 2). The range of explanations that XAI systems can provide is broad, encompassing approaches from local approximations of complex models (Ribeiro, Singh, and Guestrin 2016) to counterfactual explanations (Wachter, Mittelstadt, and Russell 2018), among others. As an illustration, take the case of LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016). LIME is a method designed to make complex ML models more understandable. It works by simplifying the model’s predictions in a specific, localized area of the data. For instance, imagine a very complicated mathematical model; instead of trying to understand the entire model at once, LIME creates a simpler, approximate model (like a straightforward linear function) that explains how the complex model behaves within a small, specific part of the data. This localized approach makes the model’s behaviour in that particular section more interpretable to humans.

A reason to develop XAI systems is that, intuitively, they are more worthy of trust than opaque systems. However, this is only the case if one accepts the premise that explanations lead to trust. Such an idea has been referred to as the *Explainability-Trust Hypothesis* (ET). Following Kästner et al. (2021),<sup>7</sup> ET could be defined as follows:

**Hypothesis 4.1.** *Explainability-Trust Hypothesis* (ET). Explainability is a suitable means for facilitating trust in a potential trustor.

---

<sup>7</sup>Definition 4.1 is a literal quote from the paper by Kästner et al. (2021), except for the words “potential trustor”, which I have used instead of the original “stakeholder”. The reason for this choice is to widen the scope of ET. This way, ET can be applied also in scenarios which involve users, policymakers, developers, etc. Besides, the following use of the term ‘means’ in this thesis is inspired by Buchholz’s (2023) means-end account of XAI.

According to this definition, the trust of a potential trustor can be fostered through explanations. ET is a very intuitive claim that has been assumed, implicitly and explicitly, by many researchers. In the *Ethics guidelines for trustworthy AI* (HLEG 2019), explicability<sup>8</sup> appears as a crucial ethical principle. Kästner et al. (2021) already provide an extensive overview of the literature on ET. Additionally, the following are some quotations that further reinforce the hypothesis::

- “We argue that explaining predictions is an important aspect in getting humans to trust and use machine learning effectively, if the explanations are faithful and intelligible” (Ribeiro, Singh, and Guestrin 2016, pp. 1135–1136).
- “[...] one of the main goals of explanation is to establish trust of people [...]” (Miller 2017, p. 58).
- “The big advantage of such systems [*explainable systems*] would include not only explainability, but [...] Most of all, this would increase acceptance and trust [...]” (Goebel et al. 2018, p. 297).
- “[...] there is a need to explain [...] so that users and decision makers can develop appropriate trust [...]” (Hoffman, Klein, and Mueller 2018, p. 197).
- “[...] the goal of the explanation is to increase the degree of trust between the user and the system, giving the user more nuance towards whether the system’s decision is likely correct, or not” (Jacovi and Goldberg 2020, p. 4199).
- “One of the main effects of the explainability of algorithms is viewed in the user’s perception of trust” (Shin 2021, p. 22).

The quotes above show that associating explainability with an increment of trust is common in the literature. However, the relationship between trust and explanations is complex.

It has been noted that ET expresses an epistemic relationship and not an empirical one (Kästner et al. 2021). This means that the focus of ET is on the subjective cognitive processes by which trust is formed, rather than on directly observable phenomena (like measurable outcomes or behaviors). Despite this being a keen observation, Kästner et al. do not develop the details of this thesis. My reading of ET is that it is not a claim connecting empirical facts. Instead, it essentially concerns knowledge: how someone can come to trust a system, or how someone can get to know that a system is trustworthy. ET tells us something about how trustors react to new information, conveyed in explanations. Thus, the focus is on the subject and their perception of the world. If they have access to explanations about how the system works, they might understand the system and therefore they might trust it. This means, that ET captures a relationship between the belief states of the trustor. That is the reason why ET is not as simple as it may seem. It says something about how a trustor comes to trust, which is a subjective process. However, ET offers an

---

<sup>8</sup>Authors such as Floridi et al. (2018) or Robbins (2019) emphasize the distinction between *explainability* and *explicability*. According to them, explicability is a wider concept which includes explainability. For the purposes of this thesis, I won’t engage with this terminological debate here.

objective<sup>9</sup> answer: explanations. This tension between subjectivity and objectivity is problematic. ET connects something objective (explanations) to something subjective (trust) without characterising the explanation recipient. This leads to a false universalization, from which certain epistemic limitations follow. I now proceed to examine these limitations.

### 4.3.2 Epistemic limitations

In the literature, it is often assumed that ET has a universal character.<sup>10</sup> There are two sides to this assumption: that every trustor reacts equally to the same explanation, and that every explanation changes the beliefs of the trustor. In this section, I show how both assumptions fail.

#### Trust and background knowledge

It is recurrent across the literature to state that explainability leads to an increment of trust in artificial systems. However, few authors (e.g. Ribera and Lapedriza (2019)) specify in which contexts this can happen. This omission contributes to the mental image of the ideal explanation recipient, who has all the previous knowledge needed to understand the explanation at hand. This way, the features and workings of explanations are discussed with not much attention paid to the target user. Guidotti et al. (2018) offer a good overview of the most popular methods for making AI explainable. The vast majority of them ignore the background knowledge of the user as a relevant factor to take into account when developing and evaluating explanations. Few studies can be found where it is specified that the explanations to which they refer are targeted at a certain demographic of the population. One of these rare examples are Ribeiro, Singh, and Guestrin (2016), who specify that only graduate students were selected for their study, precisely because previous knowledge of the field was necessary. A good explanation needs to take into account the recipient's previous understanding of the algorithms (Samek et al. 2016, p. 5), together with their needs and goals (Hoffman et al. 2018, p. 3). Background knowledge is a generally assumed requisite when talking about which kinds of explanations are effective and which are not in XAI.

Taking for granted the background knowledge of the trustor suggests by omission that every person reacts equally to the same explanation. Trust is a part of the reaction of the trustor. Thus, one could infer that a common explanation should bring about the same trust in every trustor. This is a problematic assumption that ought to be qualified. It can be the case that different trustors arrive at different

---

<sup>9</sup>Objective as opposed to subjective: an explanation is a set of sentences external to their receptor, not an internal state of them (as it is a belief, for example). Explanations are the way they are independent of the mind of the explanation recipient; in this sense, I refer to them as objective. Conversely, trust (as beliefs, perceptions or emotions) depends on the mind of the trustor, so it is subjective. I use the term 'objective' to clearly contrast it with trust's subjectivity, but properly speaking, what I mean is 'intersubjective'. For details on how I use this terminology, see section 3.2.2

<sup>10</sup>The publications cited above illustrate such an assumption. In such publications, not much else is said about how exactly explanations foster trust. Then, I understand that ET is assumed as paradigmatic with no further nuance. Additional examples of ET's assumed universality can be found in Gilpin et al. (2018), Anjomshoae et al. (2019), and Wang et al. (2019).

conclusions having received the same explanation. For example, imagine a hospital in which it is being decided whether to use an AI to diagnose pancreatic cancer. After receiving an explanation that this AI is an NN and how it operates on the patients' scans, the doctors of the hospital are convinced of its benefits, while the patients remain indifferent. The doctors had the previous knowledge required to understand the explanations, but the patients did not even know what an NN was. Not every trustor will have the same background knowledge. In the absence of such knowledge, the explanation will not alter the beliefs of the trustor and hence will be ineffective. Often, this is the case due to a mismatch between the background knowledge needed to benefit from an explanation provided by an XAI system and the background knowledge received by the actual recipient of the explanation. This mismatched is often caused by the wrong targeting of the explanation recipient. For example, if we abide by the classification by Tomsett et al. (2018), the explanations (provided by the XAI system LIME) that Ribeiro, Singh, and Guestrin (2016) discuss in their paper are targeted at operators or executors of the AI system to be explained. If the recipient of the explanation falls into a different category, such as a decision subject<sup>11</sup> the explanations will most likely not be effective. As Buchholz (2023) proposes, simpler explanations may be more effective when the goal is to foster trust, in case that the trustor's background knowledge is limited and they are not entirely familiar with the workings of AI systems.

The implementation of AI systems that fail to tailor explanations to the background knowledge of their intended audience can lead to more than just a failure in fostering trust. As Pozzi points out, the use of ML algorithms in contexts where system predictions directly affect non-experts may incur in a way of epistemic injustice (Pozzi 2023, p. 536). In her work, Pozzi focuses on medical AI, specifically in the case of Prediction Drug Monitoring Programmes (PDMPs), which are employed in US hospitals to predict a patient's likelihood of opioid addiction or misuse. These systems can silence the testimony and experiential knowledge of patients, leading to a harmful over-reliance on the AI's predictions by medical professionals. In doing so, the patients are wronged as epistemic subjects; that is, they are denied credibility, consequently affecting their own care. This constitutes a form of testimonial injustice, as the patient's testimony is unjustly discredited due to over-reliance on the system's opaque predictions (Pozzi 2023, p. 537). Moreover, I will argue that such undesirable outcomes can occur even in the context XAI. When the system is explainable, but the explanations provided fail to bring understanding to their target audience –whether due to technical complexity or mismatched epistemic assumptions– hermeneutical injustice can occur (Fricker 2007, pp. 147–152). In this case, the explanation recipient lacks the necessary conceptual tools to interpret or challenge the system's reasoning, thus remaining epistemically marginalized. Fricker's (2007) framework further enriches this argument by emphasizing that when AI explanations are overly technical, they may perpetuate an epistemic gap between the system operators and those affected by its outputs, deepening existing power imbalances. As a result, users who cannot grasp the AI's reasoning are ef-

---

<sup>11</sup>Following the terminology of Tomsett et al. (2018), a decision subject would be a user of an AI system, who does not need to be an expert.

fectively excluded from participating in informed decision-making processes, even if explanations are nominally offered.

Reflecting on the paradoxical relationship between trust and transparency pointed out by O’Neill (2002), Nguyen (2021) highlights the role of expertise in certain explanations. Nguyen argues that some explanations (or justifications) simply cannot be adapted to the general public. There are domains, to which Nguyen refers to as cognitive islands, that require a high level of expertise to fully grasp. When justifying decisions in these areas, the reasons provided by the experts are not epistemically accessible to people outside the field due to their lack of background knowledge. This does not make the justification less valid. However, the experts may be pressured to find a more accessible alternative, even if that alternative is not the real reason behind the actions that are being justified.<sup>12</sup> Nguyen’s remark affects ET directly since the idea is to use explanations as a means to increase trust. However, ineffective explanations won’t achieve this goal. In fields such as AI, finding effective explanations is challenging because, to be effective, an explanation needs to take into account the expertise of the trustor in different areas. I have shown that ET does not hold in every context. It is imperative to pay attention to whom the explanations are aimed when the purpose is to achieve their trust.

In conclusion, the effectiveness of an explanation is dependent on the background knowledge of its target, and so is the potential trust such an explanation can trigger. It is not possible to draw a single path that leads from a given explanation to trust in every trustor. Explanations contribute to trust *only* for those trustors to whom the explanations are effective. Therefore, explanations do not constitute a universal solution for the trust problem: they are a context-dependent answer to a universal question.

### Trust and bad explanations

The second ET assumption refers to the idea that every explanation changes the beliefs of the trustor. In other words, ET assumes that making AI explainable is enough to increase the trust of the trustors. However, not every explanation suffices. It can be the case that an explanation fails to trigger trust because it is a bad explanation. An explanation needs to meet certain criteria in order to qualify as a good explanation (Samek et al. 2016). Broadly speaking, I consider *good explanations* to be those that bring understanding. However, what makes a good explanation is a

---

<sup>12</sup>Nguyen illustrates this point with several examples. One particularly clear case involves a philosophy department expected to prove that certain learning outcomes have been achieved. A valid method would be for another philosophy department to assess the writing samples of the first department’s students. However, a more transparent method, understandable to the general public, might involve examining the salaries of the department’s graduates. If this latter justification is chosen, the department might replace metaphysics courses with entrepreneurship classes to boost salaries. While this shift might satisfy non-experts, it would not necessarily lead to students graduating with a deeper understanding of philosophy.

complex matter.<sup>13</sup> This discussion goes beyond the scope of this thesis, even though it is an interesting topic for further research linked to XAI and ET.<sup>14</sup>

Hoffman, Klein, and Mueller (2018) make a keen distinction between the goodness and the satisfaction of an explanation (pp. 4-5). In their words,

“While an explanation might be deemed good [...], it may at the same time not be adequate or satisfying to users-in-context. Explanation Satisfaction is defined as the degree to which users feel that they understand the AI system or process being explained to them. Compared to Goodness [*defined earlier in the paper as a property that could be evaluated a priori just by checking features such as clarity and precision*], satisfaction is a contextualized, a posteriori judgment of explanations” (p. 5, the text in italic is mine).

In this context, I find the use of the term ‘goodness’ slightly misleading. I would rather use ‘accurate’ since what is commonly understood as ‘good’ matches the definition that Hoffman, Klein, and Mueller (2018) give of ‘satisfactory’. Nevertheless, the distinction is on-target. The distinction above emphasizes the importance of the reaction of the trustor to an explanation. An explanation can be accurate (if it meets certain criteria)<sup>15</sup> but ineffective (or in the terminology of Hoffman, Klein and Mueller’s (2018), unsatisfactory) if it does not achieve its purpose. Generally, the main purpose of an explanation is to create understanding.

More specifically, in the case of AI, understanding of the system is sought for anticipation of the system’s actions, better use of the system and overall better performance –among others– (Goebel et al. 2018, p. 297). We can see then how crucial the context of an explanation is. This brings us back to section 4.3.2, where I focused on the importance of background knowledge. Other qualities, such as the clarity of the explanation, its scope<sup>16</sup> or its connection to the evidence, also play a role in the effect of the explanation on the trustor. Let me illustrate this point with an example.

Imagine we have a system that analyzes cancer data and generates visual reports to assist doctors in understanding cancer trends. A doctor asks about which cancers should be prioritised based on incidence for men and women. Imagine that the system replies as follows:

**Explanation 1.** The cancers with the highest incidence should be prioritized. You can check the number of cancer cases and deaths for different types of cancer in figure 4.1.

Explanation 1 is an accurate explanation that states true facts and it is connected to evidence. However, it would not be rare that the doctor remains confused

<sup>13</sup>Different criteria have been proposed, such as Ruben’s (1990) *particular causation*, Deutsch’s (2009) *hard-to-varyness*, or Woodward’s (2004) *manipulationism*. For a developed account of manipulationism applied to XAI, see Buijsman (2022).

<sup>14</sup>For a discussion of the quality of explanations in XAI applied to heatmaps in particular, see Samek et al. (2016).

<sup>15</sup>See Appendix A in Hoffman, Klein and Mueller’s (2018) paper.

<sup>16</sup>If the scope of an explanation is too wide, it is a bad explanation due to its imprecision. An explanation that could explain everything, does not really explain anything.

rather than being enlightened by this explanation. Just by looking at figure 4.1 it is clear that prostate and breast cancer are the ones with the highest incidence, for men and women respectively. However, it is not so clear at first sight which the third or fourth types of cancer with the highest incidence are. The response points to a confusing bar plot and it does not clearly answer the doctor’s question about which cancers should be prioritized. The doctor has to manually interpret the bars and figure out which cancers are most significant.

Alternatively, imagine that the system provides the following explanation instead:

**Explanation 2.** Figure 4.2 shows the incidence and mortality of the most common types of cancer. According to this data, prostate and lung cancer should be prioritized for men and breast and lung cancer for women, since those have the highest case and death numbers. The following types of cancer to be prioritized in men are colorectal, bladder, Non-Hodgkin’s lymphoma, kidney, pancreatic, leukemia, liver and esophageal cancer. For women, the types to be prioritized after breast and lung cancer are colorectal, Non-Hodgkin’s lymphoma, ovarian, leukemia, pancreatic, kidney, bladder, liver and esophageal cancer.

Both explanations refer to the same data. However, the second explanation presents it in a clearer and more comprehensive manner, emphasising the information the doctor is seeking. It includes an ordered list of all cancer types shown in the image, providing a complement to the visual aid and clarifying instances where the numbers are close and could be easily misinterpreted from the bars. The first explanation is likely to be ineffective, due to its lack of clarity. We can conclude that understanding is more relevant than accuracy when labelling explanations. An explanation that the explanation recipient does not understand won’t have any effect on them.

Because of this, when defining the goals of XAI, authors such as Páez (2019) propose to switch the goal from AI able to provide explanations to AI able to create understanding. In a similar line, Shin (2021) emphasises that trust is enhanced by explanations that the users can understand (and according to him, such explanations are causal ones).

In sum, the quality of the explanation needs to be taken into account when pondering its effect on the trust of the trustor. Not every kind of explanation will increase trust in every trustor. In this section, I have argued in particular that the background knowledge of the explanation recipient and the kind and quality of the explanation provided need to be taken into account to evaluate whether the explanation fosters trust.

### 4.3.3 A new version of ET

Besides being epistemologically limited, a second issue that raises problems with ET is the ambiguity about what kind of trust can be fostered with explanations. In chapter 3, I defined appropriate trust (3.6) as the kind of trust that is normatively desirable. This is the case because appropriate trust is characterized to be rooted in justified trust belief. This element of justification is crucial in cases in which the trustee is an AI system.

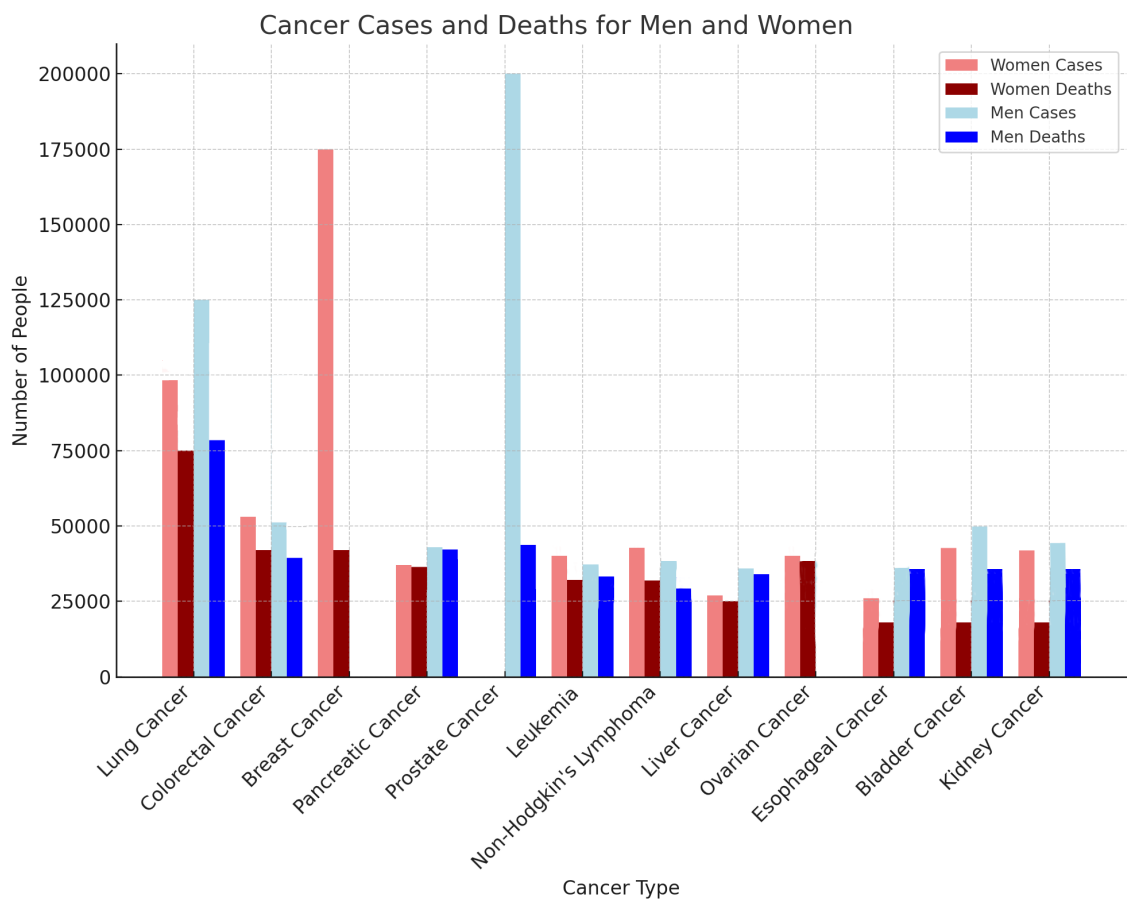


Figure 4.1: Remake of a figure that was originally published in the New York Times in 2007, using classical double column (deaths/cases) bar plot. The original figure can be found at <https://tinyurl.com/NYTfigure>.

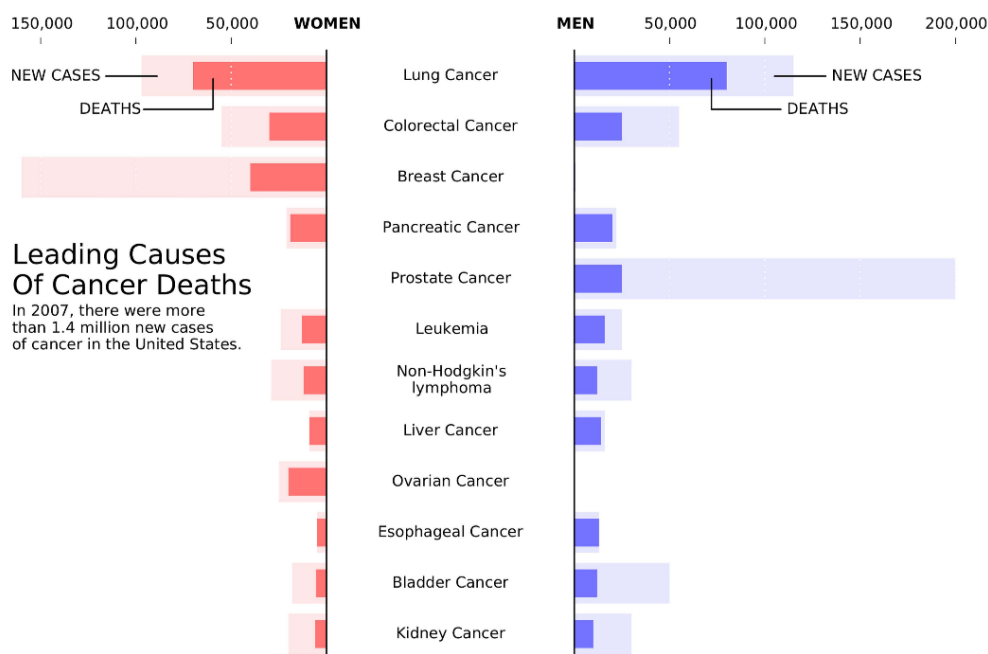


Figure 4.2: Remake of the same figure by Rougier, Droettboom, and Bourne (2014, p. 2), using four series (men deaths/cases, women deaths/cases) instead.

According to ET, providing explanations contributes to the trust-building process. Explanations would offer the trustor a pathway to understand and justify why they might believe that the trustee deserves their trust. However, it's worth noting that this relationship between explanations and trust may not always hold true in every circumstance. As I have developed along this chapter, the cases in which ET holds involve:

- (a) A trustor with the necessary background knowledge to understand the explanation.
- (b) An explanation that qualifies as a good explanation.
- (c) A system whose trustworthiness is supported by evidence, which a good explanation should reflect.

When this is the case, ET holds: explanations are a suitable means to foster a trust belief in a potential trustor. However, none of these clauses are explicit in ET. As I have shown in section 4.3.2, explanations on their own are not enough to guarantee trust. The connection between explanations and trust is contingent and dependent on the conditions above. ET describes an objective attempt to achieve trust, which is partially subjective. What I have proposed in section 4.3.2 is to shift the focus and to put care into how we approach explanations. I argue that good (or appropriate, or effective) explanations are subjective too. Taking this into account, ET could hold. But without a proper disclosure of all the necessary conditions for an explanation to be well-targeted and effective, it cannot be said that explanations lead to trust. At most, it could be said that ET is conditional on the (a)–(c) conditions

above. I lay the following question on the table: Do cases that meet the (a)–(c) conditions exist? How often do we encounter a trustor with the necessary background knowledge who receives a good explanation that properly discloses the workings of a trustworthy system? If not, ET (even dependent on the conditions above) loses its value and needs to be rejected. Given all these uncertainties and buts, I propose to reject ET and replace it with a more nuanced thesis:

**Hypothesis 4.2.** New Explainability-Trust Hypothesis ( $ET^*$ ). Explanations are a suitable means for facilitating trust in a potential trustor  $A$ , if and only if the explanation has the potential to generate in  $A$  a trust belief. To generate a trust belief, the explanation must be of good quality and tailored to  $A$ 's particular doxastic disposition. Such a disposition includes  $A$ 's background knowledge and the effect of the explanation on  $A$ .

The definition above is framed in the account of trust proposed in chapter 3. Acknowledging the doxastic nature of trust makes it clear that, not any kind of explanation has the potential to foster trust in potential trustors. To foster trust, a trust belief must be fostered. Additional conditions need to be met that have to do with the impact of the explanation on the trustor. My new thesis  $ET^*$  renders those conditions explicit which made ET fail. Thereby,  $ET^*$  describes the relationship between explanations and trust in a more accurate form than the initial thesis. The goal is to avoid problems that may arise due to oversimplification. The most salient of these problems is the failure of XAI not to take the doxastic disposition of the trustor into account. In this kind of scenario, XAI could be dismissed or unfairly criticized. A way to avoid the potential ineffectiveness of XAI (to increase trust) could be working on the kind of explanations that are offered and its suitability to the potential trustors they are aimed for. That is a possible solution that is worth working on.

---

## 4.4 Degrees of trust as degrees of trust belief

In the last section, I have pointed out the limitations of ET and proposed a more nuanced version,  $ET^*$ , which emphasises the fact that explanations can foster trust only under certain circumstances.  $ET^*$  shows that the relationship between explanations and trust is complex, and the case for artificial trustees is no exception. XAI is a wide term that refers to systems that are able to provide explanations about their workings. I label the concept as wide because many kinds of explanations can be offered, depending on the explanans and the recipient, among other factors.

Different kinds of explanations foster different degrees of trust. This variability in explanations becomes particularly advantageous for artificial trustees. This is because various AI applications require different degrees of trust from different trustors to be successfully implemented. Each application's unique demands for trust can be met more effectively when the AI provides tailored explanations suitable for each context. Let us recall a specific instance, say, case A from section 2.4.1. Alice's doctor used an NN to diagnose Alice with VS. In this case, Alice trusts the NN indirectly.

She does not necessarily require an in-depth understanding of how the system operates to place her trust in the diagnosis it provides. Instead, what matters to Alice is her doctor's trust in the system. As long as her doctor holds a justified trust belief about the system, it suffices for Alice to trust the outcome. Thus, Alice and her doctor require different degrees of trust, and therefore different explanations.

By incorporating the nuances discussed in section 4.3, ET\* matches the gradual nature of trust and sheds light on its relationship with explanations. It is uncontroversial to say that trust comes in degrees (Ferrario, Loi, and Viganò 2020). Less uncontroversial is the statement that trust requires trust belief (Holton 1994, p. 63; Kappel 2014, p. 2017). However, the doxastic nature of trust would explain its graduality since, as I will argue, different degrees of trust correspond to different degrees of trust belief. Let me go back to case A to illustrate this. As we know, Alice's doctor holds a trust belief about the NN in use; specifically, they believe that the NN is very likely to detect VS tumours accurately. Let us break down the part of the doctor's belief system that is relevant to their trust belief's coherence:

- ( $b_1$ )" The doctor believes that Alice's symptomatology matches VS'.
- ( $b_2$ )" The doctor believes that the main treatment for VS consists in extirpating it through surgery.
- ( $b_3$ )" The doctor believes that the NN that they are running through Alice's MRI scan has been trained with a varied and large enough volume of properly labelled scans.
- ( $b_4$ )" The doctor believes that the NN is able to recognize patterns and use those patterns to detect tumours such as VS in new scans.
- ( $b_5$ )" Because of ( $b_3$ )" ( $b_4$ )", the doctor believes that the NN is trustworthy: it is reasonable for them to delegate the recognition of tumours to the system.
- ( $b_6$ )" The doctor believes that, according to this NN, Alice has VS.
- ( $b_7$ )" The doctor believes that the tumour can be observed in a certain area of the MRI scan, as highlighted by Grad-CAM.
- ( $b_8$ )" Because of ( $b_5$ )" ( $b_7$ )", the doctor believes that Alice has VS.

Alice's doctor belief system<sup>17</sup> is represented by ( $b_1$ )" ( $b_8$ )" above. Their trust belief corresponds to ( $b_5$ )". This belief system is coherent since the degree to which each proposition is believed coheres with the others.<sup>18</sup> The trust belief ( $b_5$ )" is no exception: believing that the NN is trustworthy –towards some extent– is coherent with believing that the NN is able to recognise patterns ( $b_4$ )" –towards the same extent. The more entrenched the doctor's belief in the NN's capabilities, the greater

<sup>17</sup>The relevant part of it for the present argument.

<sup>18</sup>I use the concept of degrees of belief similarly to what other authors call degrees of *epistemic entrenchment*. In Gardenfors' (1988) words, "[t]his degree is not determined by how probable a belief is judged to be but rather by how important the belief is in inquiry and deliberation" (p. 17).

their degree of trust will be. Thus, understanding trust as belief-based sheds light on the trust-building process that leads to different degrees of trust.

Returning to ET\*, let us take a moment to look back at  $(b_7)$ . Grad-CAM is a popular technique used to make opaque systems explainable by highlighting the relevant areas for the system to reach an output. In case A, these areas should be the areas of the MRI where the NN found a pattern that identified as a tumour. This is the kind of visual explanation that is available to Alice’s doctor to foster their trust in the NN. Such an explanation can bolster the doctor’s confidence in the NN’s decision, fostering a greater degree of trust belief. However, it is important to note that the effectiveness of these explanations in fostering trust varies. The degree to which the doctor trusts the NN’s results depends on how well the explanation fits into their existing network of beliefs and knowledge. A more entrenched belief in the accuracy of medical imaging would make the doctor more inclined to trust the explanation provided by Grad-CAM. On the other hand, if the explanation does not align with the doctor’s background knowledge or seems overly simplistic, it might lead to a lower degree of trust. Moreover, the entrenchment of a trust belief is influenced by its resistance to counterevidence. A belief that is more resistant to counterevidence will reinforce the entrenchment of the trust belief, resulting in a stronger degree of trust. For example, if the doctor has previously encountered cases where Grad-CAM successfully highlighted the correct areas corresponding to actual tumours, this experience makes the belief in the NN’s accuracy more resistant to counterevidence. This entrenched belief would make the doctor more confident in trusting the NN’s current diagnosis, even in the face of minor anomalies or uncertainties. Conversely, if the doctor has encountered situations where Grad-CAM provided misleading or incorrect highlights, their belief in the NN’s reliability would be less resistant to counterevidence, leading to a weaker degree of trust. This variability in the entrenchment of trust beliefs highlights how different explanations can foster different levels of trust depending on how their trust belief coheres with the trustor’s belief system. In this sense, *some* explanations lead to *some* degree of trust.

---

## 4.5 The value of distrust

I opened this chapter by pointing out the value of trust. To conclude, I would like to briefly explore the value of a concept that, rather than opposed to it, appears as complementary to trust: distrust (Hawley 2014, p. 1).<sup>19</sup>

So far, I have discussed how trust in AI can be built, putting the focus on explanations. The main conclusion that can be drawn from the chapter is that *some* explanations can foster *some* degree of trust. As I argued, explainability is a good option to build trust due to its variability: different explanations can be tailored to different purposes.<sup>20</sup> Typically, these purposes correspond to the application of

---

<sup>19</sup>For details on distrust, see appendix A.

<sup>20</sup>Or to different means, if one accepts Buchholz’s (2023) means-end epistemology.

the AI; as illustrated by case A, a doctor and a patient require different kinds of explanation to build different degrees of trust in the same diagnostic AI system.

Because of the relationship between trust and explanations described in ET\*, it is commonly said that explanations bring transparency to AI, potentially open the black box in the case of opaque systems (HLEG 2019, p. 18; Gunning et al. 2019, p. 44). In section 4.4, I have attempted to analyse how explanations can not only make opaque systems more transparent but also shed light on the trust-building process. Building trust through explanations makes it easier to make explicit trust belief's justification. In case A, the doctor's trust belief  $(b_5)''$  was (partially) justified in virtue of  $(b_7)''$ . The content of the latter referred to the explanations provided by Grad-CAM, which made  $(b_3)''$ -( $b_4)''$  stronger. Being able to analyse this process helps to judge whether the trust belief is actually justified.

It could be the case that when analysing the trustor's belief system and their trust belief's coherence, one finds out that the (artificial) trustee should be actually distrusted. Once again, let us continue with the example of Alice and her doctor and add another layer to case A. In this instance, the spotlight shifts specifically to  $(b_2)''$ . Imagine that, upon further examination, the doctor realises that the area highlighted by Grad-CAM does not correspond to a tumour, but to an anomaly in Alice's jawline. This rings the doctor's alarms: as they know, the main treatment for VS consists of extirpating it through a surgical intervention that typically alters the patient's jawline. Alice has not gone through such intervention since she is in the process of determining whether she has VS. This realisation weakens  $(b_7)''$ , which leads to further investigation and revision of  $(b_3)''$ -( $b_5)''$ . It turns out, that the NN was mainly trained using images of diagnosed patients. Then, the NN focused on the aftermath of the surgery that these patients went through to extirpate the tumour, rather than on the tumour itself. This update of  $(b_7)''$  leads to its rejection and, as a consequence, the doctor's degree of belief in  $(b_3)''$ -( $b_5)''$  decreases, including the trust belief. As a result,  $(b_3)''$ -( $b_8)''$  are updated: the doctor no longer believes that Alice *has* VS, although she *may have* it. Further evidence is required, since the NN is no longer trusted enough to reach such a conclusion. Instead, NN is now distrusted. This is valuable since justified distrust can also benefit Alice's diagnosis: the doctor knows now that they need further evidence, additional tests need to be run and a second opinion may be sought.

Distrust plays a complementary role to trust in guiding trustors in their quest to identify trustworthy trustees. This process is crucial, as it helps in determining who is worthy of trust and therefore, can be relied on. Recall the definitions of **trust** and **trustworthiness** from chapter 3.<sup>21</sup> Previously, I argued that a trustee  $B$  is trustworthy to a trustor  $A$  regarding a specific task  $\delta$  if and only if it is reasonable for  $A$  to delegate  $\delta$  to  $B$ . This holds true when  $A$  has access to rational reasons that suggest  $B$  is likely to successfully perform  $\delta$ , guided by motives<sup>22</sup> that  $A$  deems appropriate. Conversely, if  $A$  encounters reasons that lead them to conclude that delegating  $\delta$  to  $B$  would be unwise,  $B$  will be perceived by  $A$  as untrustworthy. This process of discerning untrustworthiness is just as valuable –if not more so– than identify-

<sup>21</sup>Definitions 3.4 and 3.2, respectively.

<sup>22</sup>For details on my use of the term 'motive' in the case of artificial trustees, see section 2.7.1 in chapter 2, or the corresponding entry in the Glossary for a concise definition.

ing who to trust. Being able to assess both trustworthiness and untrustworthiness equips the trustor with the insight needed to make informed decisions about whom to rely on and, perhaps more crucially, whom to avoid relying on. This balancing act is essential in the AI landscape, as it ensures that trust is not blindly placed but is based on robust, verifiable attributes of the AI system. In other words, it paves the way to appropriate trust (see section 3.3.2). Building trust and distrust as complementary forces is crucial for the safe and effective adoption of AI technologies. It helps pinpoint AI systems that demonstrate reliability and integrity and avoids those that might pose risks due to flaws or biases. As discussed in chapter 3 (section 3.3), a trustee is trustworthy if it is reasonable to believe not only that they will succeed at  $\delta$ , but also that they will do so moved by motives deemed appropriate by the trustor. As shown in the example above, distrust can stem not only from unsuccessful predictions but also from unclear or unsatisfactory motives underlying those predictions. Even if a system consistently produces accurate results, its accuracy alone may not be sufficient to establish trust if that accuracy is achieved through biased methods. Likewise, if the system's predictions are accurate but the underlying motives remain opaque, this lack of transparency can still lead to distrust. Hence, the interplay of trust and distrust becomes a cornerstone in establishing a responsible and dependable AI-driven environment.



# 5

---

## *Alternatives to trust*

---

## 5.1 Introduction

When characterising human-AI relationships, reliance is often considered the primary alternative to trust. As I advanced in chapter 1,<sup>1</sup> reliance involves one party depending on another to perform a task, whereas trust can exist without actual delegation; the trustor is disposed to delegate under certain circumstances, but delegation may never occur. Reliance is conceptually less demanding than trust. It simply requires competence from the external party to perform the task. By contrast, trust involves not only competence but also the alignment of motives between the trustor and trustee. If the trustee’s motives do not align with those of the trustor, the trustor may be unwilling to delegate the task, even if the trustee is capable.

Nowadays, the term ‘trust’ is being used not only to describe relationships among humans. Both the general public and a part of academia often talk about trust and distrust in artificial intelligence (AI). In 2019, the European Commission published ethics guidelines that put the focus on the concept of trustworthy AI (HLEG 2019). The document framed trust as a necessary condition for AI’s successful implementation in society, parting from the assumption that trusting non-human entities is possible. During the last couple of years, many other documents have shared the same assumption and arrived at similar conclusions about the role of trust in human-AI relationships (Reinhardt 2022).

However, using the term ‘trust’ to refer to human-AI relationships without specifying how the term is being used leads to problematic ambiguities. How can non-human entities be the object of trust if they cannot exhibit goodwill (Baier 1986)? How is this possible for entities that cannot be held responsible for their actions (Holton 1994)? Or hold commitments (Hawley 2014)? Authors such as Nickel, Franssen, and Kroes 2010 or Ryan 2020 have pointed out the problem of using the term ‘trust’ for cases which involve AI systems as trustees. Their solution to resolve such ambiguities is to use the term ‘trust’ only to refer to relationships among humans. By contrast, when it comes to human-AI relationships, these authors propose adhering to the concept of reliance.

Reliance on technology has been conceptualised in terms of a *thin notion of trust* (Nickel, Franssen, and Kroes 2010), *rational trust* (Ryan 2020) or *computational reliabilism* (CR) (Durán and Formanek 2018). The bottom line of those approaches is the same: only humans can be trusted, whereas technological artefacts, can at best, be relied on. From that view, it follows that, when using AI, the emphasis should be placed on its technical success. By technical success, I mean that the system achieves satisfactory results most of the time in its contexts of use. For example, for a language model such as Open AI’s GPT-4, the system is considered technically successful if it is able to provide adequate answers to most<sup>2</sup> of the inputs it receives. In the remainder of this thesis, I will argue that technical success is not enough to satisfy the normative goals of AI, such as avoiding bias or respecting users’ privacy,

<sup>1</sup>For further details on the distinction between trust and reliance, see section 1.6.

<sup>2</sup>Here, I use ‘most’ in a loose way. There is no exact percentage of successful cases from which a system is considered technically successful. What is considered a high rate of success varies depending on the system to be rated and its application.

for example. Therefore, reliance falls short in capturing the normative relationship between humans and artificial systems.

In this chapter, I will analyse the position of authors who claim that only humans can be trusted and that AI, at best, can be relied on. To illustrate this position, I have chosen papers by Nickel, Franssen, and Kroes (2010), Ryan (2020), and Durán and Jongsma (2021). In sections 5.2, 5.3 and 5.4, I will provide an overview of each paper, followed by a summary of the points I agree with and my critiques. In section 5.5, I will summarise the reasons for my disagreement with the authors above. This will serve as a preparation for my argument in favour of trust as a suitable goal for human-AI relationships, which I will develop in the next chapter.

---

## 5.2 Nickel, Franssen and Kroes: a thin notion of trust

I would like to start this section with a paper by Nickel, Franssen, and Kroes titled “Can We Make Sense of the Notion of Trustworthy Technology?”. The authors do not specifically focus on AI, but technical artefacts in general. However, I consider their notion of *thin trust* to be a precursor to some of the views on trust in AI that emerged over the following decade.

### The account

Nickel et al. consider both trust and trustworthiness as three-place relations “between a trusting agent  $A$ , some trusted entity  $B$  and an anticipated, desirable performance  $x$ ” (Nickel, Franssen, and Kroes 2010, p. 431). This way, trust refers to the relationship between  $A$  and  $B$  regarding  $x$ , in which  $B$  can be considered trustworthy by  $A$  with regard to the performance of  $x$ . Building on this basis, the authors distinguish between two different accounts of trust:

- The *pure rational-choice account*. Trust is described as a decision made by the trustor after evaluating the potential risks and rewards of depending on the trustee, along with considering relevant alternatives, ultimately concluding that it is beneficial to rely on the trustee’s performance.<sup>3</sup>
- The *motivation-attributing account*. Trust is conceived as an attitude of the trustor towards the trustee in which the former expects the latter to behave in a certain way, moved by certain motives. Such motives are usually related to some moral obligation.

Nickel et al. connect motive-attribution accounts of trust and moral obligation since, often, the motives attributed to the trustee point towards moral integrity (McLeod 2021) or goodwill (Jones 1996). Thus, this kind of accounts<sup>4</sup> capture an essential feature of interpersonal trust, but, according to the authors, cannot be

---

<sup>3</sup>Nickel, Franssen, and Kroes draw from the framework provided by Coleman (1990, pp. 91–114).

<sup>4</sup>Into which my BBD account falls.

transferred to trust in technology since only humans can have mental states and, therefore, motives.

Given the limitations of the motive-attribution accounts of trust, Nickel et al. resort to pure rational-choice accounts and propose to understand the relationship between humans and technological artefacts not in terms of trust, but in terms of reliance. Building on an engineer-like notion of reliability, they characterise reliance as the expectation of a desired behaviour, defining it as the ability to “*remain functional*” (Nickel, Franssen, and Kroes 2010, pp. 433-34). Whether someone or something is reliable or not can be measured quantitatively, given the probability that they will perform as expected.

The authors go a step further and concede that there are artefacts that exhibit agent-like behaviour to which it may be attributed a derived sense of trust from their designers. That is, the artefacts are indirectly trusted, but properly speaking, the trustee is the person responsible for such artefacts. Nickel et al. call this a *thin notion of trust* (Nickel, Franssen, and Kroes 2010, pp. 439-40). According to them, a thin notion of trust can be attributed to technical artefacts that seem to adapt to the user’s values that are considered relevant for the situation in which the artefact is being used, like, for example, a car navigation system. This thin notion of trust derives from the artefacts’ designers, rather than trust in the artefacts themselves. The reason is that the designers, developers and deployers are the ones who have interests and motives that can be in alignment with the users’ interests.

### Points of agreement

In their paper, Nickel et al. describe full-blown trust as motives-based, meaning that it requires that the interest of the trustor and the trustee are in alignment (Nickel, Franssen, and Kroes 2010, p. 443). I agree with them on this point and I also consider motives to be the key difference between trust and reliance (for further details, see section 1.6 in chapter 1 and section 2.7 in chapter 2).

The authors establish a connection between trust and moral responsibility, even though, according to them, it does not apply to trust in technology (which is actually reliance) (Nickel, Franssen, and Kroes 2010, p. 443). My view on trust is also linked to moral responsibility, as I will develop further in the next chapter. In contrast with Nickel et al., I will argue that this is the case both for interpersonal trust and trust in AI.

Nickel et al. make a clear distinction between trusting the designers of an artefact and trusting the artefact itself (Nickel, Franssen, and Kroes 2010, p. 238). I agree that these are distinct concepts that should not be confused. While the users’ trust in the designers (and the developers and deployers) can be better understood as institutional trust, trust in the artefacts falls under what I have characterised as trust in AI. It is important to note that trust in AI is a complex kind of trust that involves not only trust in the system but also in the humans responsible for the system. I will talk about this in the next chapter in further detail. The basis for this can be found in section 2.3 in chapter 2.

Finally, an important point in Nickel et al.’s paper is that artefacts that exhibit agent-like behaviour pose a special case (Nickel, Franssen, and Kroes 2010, pp. 439–

40). I agree on this, even though I understand such a special case differently. Again, further details can be found in section 2.3 in chapter 2.

### Critique

The main takeaway from Nickel et al.’s paper is that trust is motives-based and, since technological artefacts cannot hold motives, full-blown trust is exclusively human. I agree with the first part of this statement, as my own account of trust is also motives-based (see chapter 3). However, whether the second part of their conclusion holds depends on the interpretation of motives. In the paper, the authors reiterate that motives are encapsulated in mental states (Nickel, Franssen, and Kroes 2010, p. 433). My disagreement begins with this starting point. It is relatively uncontroversial to define motives as the combination of desire and belief that drives an agent’s actions (Davidson 1963, p. 687). This definition fits Nickel et al.’s framework. Aligned with their view, motives are considered exclusively human, as AI systems lack the capacity to hold beliefs or have desires. This is a common critique against the concept of trust in AI. However, in relation to AI, I adopt a broader definition of motives, considering them as the criteria that favour a specific course of action.<sup>5</sup> If we interpret motives in AI systems in the analogous manner I propose, Nickel et al.’s conclusion does not hold: AI systems can be seen as having motives too (even if they are merely *analogous* to human motives rather than identical). It is fair to say that their understanding of motives is widely accepted and that my use of the term is by no means canonical. However, as AI is a novel technology, society is still learning how to interact with it. We have yet to establish undisputed social codes of conduct or a widely shared vocabulary for these interactions. This evolving mindscape is still under construction. It is not uncommon to describe AI systems using terms analogous to typical human behaviour, such as ‘learning’, ‘recognition’, ‘perception’ or ‘decision making’. Therefore, I believe that drawing an analogy between human motives and motives understood in this wider sense is fitting in the context of AI-human interaction.

The second point of my critique addresses the concept of *thin notion of trust*, as derived from interpersonal trust (in the people responsible for the artefacts). I agree with Nickel et al. that this concept of derived trust applies to artefacts exhibiting agent-like behaviour (noting that their paper was published in 2010 and did not specifically focus on AI systems). However, I will argue that AI systems take this a step further. As previously stated,<sup>6</sup> AI is widely considered not merely a technological artefact but a socio-technical tool (STT) (Rieder, Simon, and Wong 2021; Ananny 2016; Benk et al. 2022; Jones, Artikis, and Pitt 2013). Partly, this means that, unlike mere tools, the use of AI systems goes beyond extending a particular user’s ability. AI systems *seemingly* ‘master’ abilities themselves, conducting apparent decision-making. By this, I do not mean that AI systems have (full) agency. In any case, some sort of *quasi-agency* (Coeckelbergh 2019) could be attributed, more as a result of how users engage with the systems than of the systems’ abilities per se.<sup>7</sup> This attributed *quasi-agency* goes beyond the ‘agent-like behaviour’ exhibited by artefacts

<sup>5</sup>Further details can be found in chapter 2, section 2.7

<sup>6</sup>See section 2.3

<sup>7</sup>Further details can be found in chapter 2, section 2.3.2

such as car navigation systems. In the case of AI *quasi-agents*, their appearance of agency –or *quasi-agency*– leads to a stronger inclination to anthropomorphize some AI systems and interact with them as we do with other humans (Placani 2024). It could be argued that a similar phenomenon occurs with the kind of artefacts that Nickel et al. reference, as evidenced by people yelling at their TomTom during a traffic jam. However, the level of intuitive anthropomorphism, and the accompanying attributed agency, is much more pronounced with AI systems. This is partly because users do not perceive the underlying mechanisms that enable the system to mimic human skills, seeing only the manifestations of such replicated skills. AI introduces a new level of complexity, resulting in an apparent autonomy that significantly differs from that of simpler, more predictable artefacts. Even if users are familiar with how AI functions, they still interact with an artefact that mimics human behaviours. Consequently, the natural response is to react to these behaviours as one would with other humans. This dynamic of mimicry and engagement with mimicked behaviours includes trusting relationships. In conclusion, while Nickel et al.’s thin notion of trust applies to artefacts that display agent-like behaviour, AI systems do not fall within this category. Instead, I regard them as *quasi-agents*. Quasi-agents mimic human agency further than the artefacts that display agent-like behaviour that Nickel et al. refer to, since quasi-agents can exhibit unpredictable behaviour. This is the case for some AI systems, whose outcomes cannot even be predicted by their designers. <sup>8</sup>

Finally, the last section of Nickel et. al’s paper addresses the topic of trust in socio-technical systems.<sup>9</sup> They start the section by stating that “[i]f trust requires the taking into account of interests and the possibility of a conflict of interests, then trust in technology must be derived in some way from *interpersonal trust*” (Nickel, Franssen, and Kroes (2010, p. 441), emphasis mine). From that premise, they argue that in hybrid systems involving both people and technical devices, the user’s actions are constrained by external rules and the actions of the other agents participating in the system. Using this kind of system, the user is not receiving just a product but a *service*. Therefore, they conclude that “we trust the designer/manufacturer of a product, but not the product” (Nickel, Franssen, and Kroes 2010, p. 441). Later in the section, they discuss the role of the operators of socio-technical systems as trustees. According to the authors, socio-technical systems entail a net of promises, motives and responsibilities in which the user’s trust is distributed in degrees among the different people who participate in the system. But according to Nickel et al., technical components of the system are not trusted at all: they are merely relied on. In other words, only the human components of a socio-technical system can be objects of trust, but not the technical artefacts. In line with their previous argument, the reason is that only humans can have mental states that encapsulate interests and motives. Therefore, only humans can be trustees with motives in alignment with the motives of the trustor. I have already argued for my disagreement with this notion of motives when it comes to technology. But in addition, it is worth addressing the

---

<sup>8</sup>To clarify: I do not suggest that every AI behaviour is unpredictable. Only that, on some occasions, it can be (whereas, in principle, the artefacts referenced by Nickel et al. are never unpredictable).

<sup>9</sup>In the widespread use of the term, not as in STTs. For details on this distinction, see section 2.3 in chapter 2.

premise from which their argument on socio-technical systems starts. In this last section, Nickel et. al aim to draw a parallel between interpersonal trust and trust in socio-technical systems. The whole section seems to ignore institutional trust and leaves the reader wondering why not to establish this parallel instead. The idea that socio-technical systems incorporate networks of promises is similar to the discussion about institutional trust in the literature (Pettit [1995]; Lahno [2001]), rather than interpersonal accounts of trust. In accounts of trust that focus on interpersonal relationships, it is discussed the role of an individual trustee. By contrast, institutional accounts explore the role of trustees composed by several people who form a network of promises. In most accounts of institutional trust, it is considered that the members of an institution are trusting in virtue of their membership. Then, an individual can trust the Police, as an institution, and because of that also trust individual police officers. This idea seems in line with their notion of distributed trust among developers and operators. More specifically, Nickel et. al point out that there are different kinds of operators (higher or lower level) and that trust can be distributed among them according to their status. If one adopts the notion of motives that I proposed, this very same argument can be extended to (some) of the technical parts of socio-technical systems: the ones that are able to have motives (in this wider sense). As discussed in chapter 2 (section 2.7), AI systems can fall into the latter category. Then, the idea is that AI is trusted as a socio-technical system, meaning that different levels of trust are established with different system parts. For example, the trustor can trust both the operator of the AI system and the AI system itself. The trustor trusts the operator if the trustor believes that the operator will do their job correctly, moved by appropriate motives, such as a passion for their job. The trustor trusts the AI system if the trustor believes that the system will provide accurate results, product of an appropriate working process. What the trustor considers to be an appropriate process for the AI system to follow will depend on the specific AI system and the task that such a system is meant to perform. But just in very general term, let us say that an appropriate process for an image classification system would include to focus on the relevant pixels of image for its classification, instead of relying in some other correlation factors, for example. In sum, trust in socio-technical systems can be distributed among different kinds of trustees such as developers, operators, policy makers, stakeholders and socio-technical tools as AI systems.<sup>10</sup> This does not mean trust in AI is identical to trust in people or institutions; in fact, I conjecture the former involves a new kind of trust. Because this new kind of trust is analogous to more paradigmatic forms of trust, but clearly distinct from them, the label ‘trust in AI’ becomes necessary.

---

<sup>10</sup>Again, for details on the distinction between socio-technical systems and socio-technical tools, see section 2.3 in chapter 2.

### 5.3 Ryan: rational trust

In this section, I turn to an examination of a paper by Ryan (2020): “In AI We Trust: Ethics, Artificial Intelligence, and Reliability”. As the title anticipates, this paper also resorts to a notion of reliance in order to address the possibility of trusting AI. In particular, Ryan argues that the only kind of ‘trust’ that is possible to place in AI systems is *rational trust*, which is actually a form of reliance.

#### The account

In his paper, Ryan argues that AI does not have the capacity to be trusted. His main argument is that trust requires the attribution of affective or normative motives to act. Affective motives primarily refer to goodwill guiding the trustee’s actions towards the trustor. Normative motives refer to the trustee’s tendency to act according to what they should do (Ryan 2020, pp. 4–5). Since AI neither possesses emotive states such as goodwill nor can be held responsible for its actions, AI cannot be the object of trust (Ryan 2020, p. 2). Ryan describes trust as a phenomenon involving a trustor  $A$ , a trustee  $B$  and some action  $X$ . Influenced by the literature on interpersonal trust, he characterises trust as follows:

- (i)  $A$  has confidence in  $B$  to do  $X$ .
- (ii)  $A$  believes  $B$  is competent to do  $X$ .
- (iii)  $A$  is vulnerable to the actions of  $B$ .
- (iv) If  $B$  does not do  $X$  then  $A$  may feel betrayed.
- (v)  $A$  thinks that  $B$  will do  $X$ , motivated by one of the following reasons:
  - (v.i) Their motivation does not matter (rational trust).
  - (v.ii)  $B$ ’s actions are based on a goodwill towards  $A$  (affective trust).
  - (v.iii)  $B$  has a normative commitment to the relationship with  $A$  (normative trust) (Ryan 2020, p. 5).

Options (v.i), (v.ii) and (v.iii) refer to different accounts of interpersonal trust, which Ryan labels in parentheses. As pointed out above, AI lacks emotive states and it can not be held responsible for its actions. Therefore, Ryan concludes that *rational trust* is the only kind of trust that can be legitimately placed on AI. Rational trust refers to the kind of trust in which the trustor makes themselves dependent on the trustee concerning some specific action, regardless of the trustee’s motivation to execute such action. This concept is equivalent to reliance (Nickel, Franssen, and Kroes 2010; Hawley 2014). Consequently, Ryan concludes that it is possible for humans to rely on AI, but not to trust it.

### Points of agreement

In his paper, Ryan presents trust as a moral concept (Ryan 2020, p. 2). In the next chapter, I will provide my reasons for agreeing with him on this.

Ryan provides a definition of interpersonal trust that aims to capture the most popular accounts of trust, depending on the role considered for the trustee's motives in their relationship with the trustor (Ryan 2020, p. 5). This goal overlaps with section 2.5, in chapter 2 of this thesis. There, I also sketched an overview of the most relevant accounts of interpersonal trust, with the aim of better understanding trust in humans. Like Ryan, I also deemed the motive-based accounts the most convincing ones, incorporating this trait in my own account of trust.

Ryan distinguishes between reliance and trust by stating that, while the former is only based on past performance, the latter requires additional conditions to materialise (Ryan 2020, p. 11). His distinction is similar to my own, which can be found in sections 1.6 in chapter 1 and 2.7 in chapter 2.

Ryan points out that the organisations that develop AI have (part of the) responsibility for the outcomes of the systems they release. Such a responsibility should be acknowledged, even if it comes at the cost of a slower pace in the development of the AI industry (Ryan 2020, p. 14). I strongly agree with this, and I will argue for it in the next chapter.

In line with the previous point, and in contrast with Nickel, Franssen, and Kroes (2010), Ryan establishes the parallel between trust in AI and institutional trust (rather than interpersonal) (Ryan 2020, pp. 15–16). This is a valuable parallel that explains why attempts to fit trust in AI into accounts of interpersonal trust often seem forced. As I argued in chapter 2,<sup>11</sup> trust in AI shares common traits with both interpersonal and institutional trust. Because of that, it deserves to be labelled as trust. However, since there is not a full overlap with either interpersonal or institutional trust, trust in AI can be better understood by being conceptualised as a category of its own.

### Critique

While I agree with Ryan that rational trust is actually mere reliance, I do not believe that this is the only kind of 'trust' that can be placed in AI. The core of my disagreement lies in his characterisation of trust, particularly in point (v). From (v) it follows that the trustor trusts the trustee when the former thinks that the latter will perform some action  $X$  motivated by affective or normative reasons. I beg to disagree. I proceed to show why the attribution of neither affective nor normative motives is strictly necessary for trust to develop.

Ryan's allusion to affective motives is inspired by philosophers such as Anette Baier and Karen Jones, who hold views on trust revolving around the concept of goodwill. In her foundational paper, Baier (1986) characterizes trust as an attitude of accepted vulnerability. A decade later, Jones (1996) redefined trust as an affective attitude. According to her, the trustor trusts the trustee when the former perceives the latter both as competent and good-willed. That is, the trustee normally acts as the trustor expects them to act because the trustee knows they are being counted

---

<sup>11</sup>See sections 2.4 and 2.7.

on (Jones [1996](#), p. 8). On their part, the trustor trusts because they think that the trustee cares about the trustor and aims not to let the trustor down. Both Baier and Jones propose accounts which assume an emotional connection between the trustor and the trustee in virtue of which the latter cares about not letting the former down.

While I share Baier and Jones' idea that the trustee's motivation to act plays a key role in trust, I will argue that such motivation does not need to have an affective or emotional character. Often, trust is paired with an emotional relationship between the trustor and the trustee. Therefore, the focus that a part of the literature on interpersonal trust has put on the trustee's goodwill is understandable. However, there are trusting relationships that involve less emotion than what Baier and Jones sketch. Most of the time, trustors base their trust on past experiences and the nature of such experiences is varied. Sometimes, we trust others based on their professionalism or expertise, rather than their goodwill. Examples of this can be found in employer-employee or doctor-patient relationships. I like to think that my doctor cares about my well-being and, because they care about me, they will do their best to help me. In other words, I like to think that my doctor has goodwill towards me. However, I do not know that. All I know is that my doctor's job is to take care of my health the best they can. They may find me irritating, unpleasant and have no positive feelings towards me at all. However, the assumption that my doctor's professionalism is above their personal judgments is enough for me to trust them (O'Neill [2002](#), p. 14). We trust judges, civil servants, police officers, etc. Not all of these trustees necessarily *care* whether trustors are let down. However, it is important to notice that the trustees do have motives to not let trustors down, even in the absence of goodwill. A judge may not have personal concern for me, but there can be other motives that lead them to deliver a fair judgment. For instance, they might be motivated by a desire to cultivate a reputation as a fair person, which indirectly influences the outcome of my case (positively). As long as the judge has reasons to uphold justice in my case, I can trust their decision-making. In sum, my only remark in this regard is that the nature of such motives does not need to be affective.

Following Ryan ([2020](#)), when characterising trust, another possibility is to focus on the trustor's attribution of normative motives to the trustee. This kind of account goes a step further with respect to the affective motives summarized above. Normative accounts of trust state that the trustor does not only rely on the trustee's goodwill, but they feel that the trustee owes them such goodwill (Cogley [2012](#), p. 31). According to this view, the trustor does not trust the trustee with whatever action  $X$ .<sup>12</sup>  $X$  must be moral, and the trustee should be able to recognise it as such. Thus, qualifying as a trustee implies holding moral agency and, consequently, being subject to potential blame (Ryan [2020](#), p. 13).

In a trusting relationship, the trustor believes that the trustee will adhere to some sort of norm that the two of them share. For example, when I trust my neighbour to take care of my cat, I do it because I believe that both of us agree that my

---

<sup>12</sup>Cogley ([2012](#)) coins this restriction in order to avoid what he calls 'the trickster problem'. This problem refers to situations in which a party relies on another's goodwill to pursue something against the good-willed party's interest. Typical examples comprise scammers, criminals and, in general, anyone who takes advantage of a good-willed victim.

cat's wellbeing is important and that looking out for the animal is the right thing to do. There is some sort of *internal normativity* that applies to the members of the trusting relationship. However, the term 'normative motives' refers to the motives behind actions that are based on adherence to certain norms that a wider social group shares. In trusting relationships, the trustor and the trustee do not necessarily adhere to moral or social norms, but to the norms that both of them implicitly agree on. The trustee behaves how both the trustor and themselves believe they should behave. However, an external party may very well disagree. For example, members of a criminal organization can trust each other to perpetuate crime. Criminals who trust each other are motivated to act according to their shared values and norms. However, these motives are hardly normative since the norms to which criminals adhere are shared by too few people to be proper norms. Because of this, even if trust implies some sort of internal normativity, the above example involving trust without normative motives shows that we better avoid talking about normative motives in general.

In relation to the trustor's attribution of normative motives to the trustee, Ryan introduces the idea that the trustee needs to be a full moral agent (p. 10). One of his main arguments that he proposes is that, if a party can not be held responsible for their actions, then they do not qualify as a trustee.<sup>13</sup> However, there are cases in which the responsibility *should not* be traced to the trustee since they could never be fully responsible. I proceed to explain why.

When we trust, we assume the risk of things not going as we want or expect. Such a risk can materialise due to a variety of reasons, and not all of them can be attributed to the trustee. For example, if I trust my neighbour to keep my cat safe while I'm on holiday, I am aware that a series of unfortunate events could happen to my cat over which my neighbour has no control. My cat could eat a bee, have a heart attack or jump through the window, despite my neighbour's desperate efforts to chase after the animal. If any of that happened, I would not blame my neighbour, or at least not fully. Trust implies moral responsibility. However, I disagree with Ryan on how such a responsibility could be potentially distributed in different kinds of situations. Trust is a complex phenomenon and the responsibility for the trusted action *X* does not necessarily fall completely over the trustee *B*. Because of this, I do not consider that the trustee needs to be a full moral agent since the full responsibility for *X* does not necessarily lie with them.

In sum, my critique of Ryan's view lies in his characterization of trust. If one understands trust as Ryan characterises it –a relationship between two parties in which one attributes affective or normative motives to the other–, then the conclusion that AI can not be trusted does indeed follow. But this disjunctive characterization is not necessary, as I will show. But before that, let me examine another view that leans towards reliance when it comes to human-AI relationship.

---

<sup>13</sup>However, that does not mean that the trustee is always *fully* responsible for the consequences that follow from performing (or failing to perform) the action *X* with which they have been trusted by the trustor. Adopting such a position would be oversimplistic and Ryan does not do so. He states that trustees can only be full moral agents, but he does not make the explicit jump to the conclusion that trustees are always fully responsible for the outcome of the actions they are trusted with.

---

## 5.4 Durán and Jongsma: computational reliabilism

In this last section, I'd like to address the concept of *computational reliabilism* (CR) as an alternative to trust in AI. Even though the concept was first introduced by Durán and Formanek (2018), I will focus on a later paper by Durán and Jongsma (2021) in which they revisit CR and apply it to the context of medical AI.

### The account

In their paper, Durán and Jongsma (2021) address the black box problem, discussing how the epistemic opacity inherent in some AI systems gives rise to ethical concerns that undermine trust. These issues primarily involve accountability, transparency, discrimination, and bias. Durán and Jongsma particularly focus on transparency, which is commonly seen as a way to justify the reliability of AI systems. In Durán and Jongsma's (2021) words, transparency is an “epistemic manoeuvre intended to offer reasons to believe that certain algorithmic procedures render a reliable output” (p. 330). Thus, transparency is widely seen as a desirable feature for AI systems. However, Durán and Jongsma claim that transparency does not solve the opacity problem. Their argument is that, to make an AI system transparent, another AI system is needed to provide explanations. They argue that this second system is as opaque as the first one was, before explanations were provided. In XAI, the opacity of the system that makes the original system transparent is accepted. Therefore, xAI does not solve the problem of opacity but transfers it from one system to another. Quoting Durán and Jongsma:

“[...] transparency is a methodology that does not offer sufficient reasons to believe that we can reliably trust black box algorithms. At best, transparency contributes to building trust in the algorithms and their outcomes, but it would be a mistake to consider it as a solution to overcome opacity altogether” (Durán and Jongsma 2021, p. 330).

In response to the problem of opacity, Durán and Jongsma propose *computational reliabilism* (CR). CR accepts opacity rather than attempting to eliminate it, positing that transparency is not necessary for establishing trust. CR does not require transparency to achieve trust because, according to Durán and Jongsma, reliance suffices in the case of AI. They define CR as a framework where “researchers are justified in believing the results of AI systems because there is a reliable process (ie, the algorithm) that yields, most of the time, trustworthy results” (Durán and Jongsma 2021, p. 332). This means that if a system provides accurate results most of the times, it is considered to be reliable and its results can be trusted. Additionally, they outline several indicators of reliability that must be met, including “verification and validation methods, robustness analysis, a history of (un)successful implementations, and expert knowledge” (Durán and Jongsma 2021, p. 332). These criteria are designed to evaluate the accuracy of the outcomes, including expert knowledge. In the paper, the focus is on medical diagnosis, with human experts having the final say in validating the correctness of a diagnosis. The paper does not elaborate extensively

on the criteria used by human experts to make these final decisions, leading to the interpretation that the primary concern is the accuracy of the system's results.<sup>14</sup>

### Points of agreement

According to Durán and Jongsma, the black box problem is not necessarily solved by providing explanations for the systems' outcomes (Durán and Jongsma 2021, p. 330). I agree with this, even though my reasons are different than the ones presented in their paper. In chapter 4 (section 4.3) I argued that merely providing explanations does not necessarily build trust in the explanation recipient. Unlike Durán and Jongsma, I focused on the need for good explanations that are tailored to the recipient's background knowledge.

In the paper, the authors argue that in fields in which AI is used in high-stakes scenarios, such as the medical field, AI systems can provide recommendations, but human experts are required for final decisions (Durán and Jongsma 2021, p. 332). I agree with them about this. This view goes in line with the conceptualization of AI as STTs, rather than mere artefacts (for details, see section 2.3 in chapter 2).

### Critique

Durán and Jongsma present CR as an alternative to transparency to achieve trust in (medical) AI systems. According to them, working with XAI offers transparency, but transparency does not solve the black box problem: instead, the opacity problem is transferred from the original opaque system to the XAI system meant to explain the outcomes of the original system. Thus, CR offers a better route to trust in AI, which ultimately should depend on how likely the system is to provide accurate outcomes. In other words, in the CR framework, trust in AI equals reliance.

I have several points of disagreement with the argument above. The first one concerns the idea that XAI systems merely transfer the opacity problem from one system to another one. Durán and Jongsma argue that the opacity is merely transferred to the XAI system. However, the level of transparency achieved with XAI is fundamentally different from the initial opacity. The explanations provided by XAI, even if not perfectly clear, offer a bridge between complex algorithmic decisions and human understanding, which is a significant improvement over complete opacity (Xiong et al. 2024, pp. 12–13; Adadi and Berrada 2018, pp. 52147–48). Even if XAI introduces another layer of complexity, this layer is typically more accessible and understandable than the original black box. XAI often provides explanations in terms that are more familiar and interpretable to humans (e.g., feature importance, decision trees, and rule-based summaries). Durán and Jongsma might see this as shifting complexity. However, making progress towards transparency, even if incremental, is valuable. It is better to have some level of interpretability, which XAI provides, than to have none at all.

Secondly, Durán and Jongsma's acceptance of opacity leads them to defend trust in black boxes. That is, according to them, it is possible to trust AI systems

---

<sup>14</sup>For further details on the role of experts, see Durán and Jongsma (2021, p. 332) and Durán and Formanek (2018, pp. 662–63).

solely<sup>15</sup> based on their performance, ignoring the process that the systems follow to perform that way and, therefore, ignoring the systems' motives<sup>16</sup> to act. I have already argued for my view on trust as motives-based in sections 2.7 (chapter 2) and 3.2.2 (chapter 3). As a reminder, the core of my argument lies in the need to avoid trusting trustees that *happen to* behave as the trustor wants, but are moved by motives that the trustor does not approve of. When the latter is the case, we have reliance but not trust. This fits the CR framework, which equates trust in AI with reliance. However, I will argue that relying on AI is not enough to satisfy the moral demands of some of the high-stakes scenarios in which AI systems are used. I have given some hints about why this is not desirable in section ?? (chapter 2), but I will develop my argument in full in the next chapter. For now, let us just say that when the stakes are high, the trustor should be able to justify their trust in case things go wrong. The trustee's motives to act play a key role when justifying the trustor's potential delegation in the trustee. Without taking the trustee's motives into account, the trustor takes the risk of relying on correlation rather than on justified predictions. Consider case A from section 2.4.1 (chapter 2). Without Grad-CAM making the diagnostic system more transparent, there is a higher risk of reaching a wrong diagnosis. The reason is that until encountering an inaccurate prediction, the system could be trusted based on results that happened to be correct due to an unrelated correlation.<sup>17</sup> Therefore, accuracy is not enough to justify trust: the motives behind the system's accurate results should be taken into account to deem the system (un)trustworthy. A final thought on this second disagreement concerns Durán and Jongsma's reliability indicators (Durán and Jongsma 2021, p. 332; Durán and Formanek 2018, pp. 656–58). Drawing from Durán and Formanek's CR framework, Durán and Jongsma (2021) quote the following indicators as necessary steps to judge a system's reliability: "verification and validation methods, robustness analysis, a history of (un)successful implementations, and expert knowledge" (p. 332). The discussion of these indicators introduces a valuable nuance in CR, namely, that the likelihood of a system providing accurate results is not enough to evaluate its trustworthiness (or in their terms, whether the system is computationally reliable). Nevertheless, the four indicators mentioned above still focus heavily on accuracy, failing to address the motives behind a system's performance. Interestingly, the authors put an emphasis on expert knowledge, which may open the door to (the expert) addressing additional criteria such as motives. However, this is my own speculation.

Finally, the last of my concerns with CR is its inability to clarify wrong performance a posteriori. Often, AI is used in contexts in which the reasons behind a mistake need to be understood to avoid further incidents and to justify the failure the affected parties. Adadi and Berrada (2018) expose the case of a self-driving car

---

<sup>15</sup>To be precise, Durán and Jongsma also introduce the concept of reliability indicators, which are further criteria to trust AI systems (beyond their predictions). I will cover this point shortly.

<sup>16</sup>Here, I stick to the concept of motives that I have presented in this thesis. Further details can be found in section 2.7 in chapter 2.

<sup>17</sup>Such as the diagnostic system having been trained mainly on scans of patients who had already gone through surgery to extirpate the VS tumour, and the system focusing on the aftermath of the surgery rather than on the tumour itself. For the full version of this example, see section 4.5 (chapter 4).

that killed a woman in Arizona. It was necessary to determine whether the accident occurred due to the car's software wrongly identifying the woman as an inanimate object or due to any other reason. If the causes of this kind of accident remain unknown, it would not be responsible to continue to use the systems that cause them since the flaw in the system will persist. In Adadi and Berrada's words, "[o]nly an explainable system can clarify the ambiguous circumstances of such situation and eventually prevent it from happening" (p. 52143).

---

## 5.5 Recapitulation

In this chapter, I have examined three influential papers in which it is argued that trust is not a suitable concept to describe human-AI relationships. Instead, different forms of reliance are proposed. I have explained the different reasons why I disagree with these approaches, which I proceed to recall as take-away:

- Aiming for reliance as a desirable human-AI relationship fails to acknowledge the role of motives when deeming whether a task should be delegated to an AI system.
- Denying the concept of trust in AI by arguing that only agents can be trusted reduces the concept of trust to interpersonal relationships. This argument ignores the theory on institutional trust, which is vast and influential.<sup>18</sup>
- In line with the point above, advocating for reliance in AI as an alternative to trust often reduces AI systems to mere artefacts, denying their socio-technical dimension.
- Finally, describing trust in AI as reliance often comes from the baseline that trusting relationships are very nuanced while human-AI relationships should be based on accuracy only. This is a reductive view on trust in AI against which I have argued at different points of this thesis, but most importantly in chapter [3](#).

In the next chapter, I proceed to extend my critique on why reliance does not suffice to describe some of the appropriate human-AI relationships that are already taking place. To do so, I will explore the moral dimension of trust in AI.

---

<sup>18</sup>See section [2.6](#) in chapter [2](#).



# 6

---

## *The moral component of trust*

---

---

## 6.1 Introduction

Society is built upon the interactions of its members and the relationships that form among them. There exists a wide variety of relationships that can be established between different people, such as friendship, respect, or submission. In some of these relationships, one party expects the other(s) to behave in such a way that leads to morally significant outcomes.<sup>1</sup> When that is the case, there is moral responsibility. A good example is the doctor-patient relationship. Once two people engage in this kind of relationship, the one who occupies the role of the doctor has the moral responsibility not to harm the patient. Similarly, when a trustor and a trustee become part of a trusting relationship, the trustee should abide by their responsibility of fulfilling the trustor's expectations.

So far in this thesis, I have explored the nature of trusting relationships, human-AI relationships, and whether the latter can be understood as instances of trust. Given the significant similarities and shared expectations between the two, the answer seems to be affirmative. The final question this chapter addresses is how the moral dimensions of trusting relationships impact those involving AI as trustees.

Unlike reliance, I will argue that trust is a relational concept that implies moral responsibility. Trust refers to relationships where a trustor willingly makes themselves vulnerable to a trustee and accepts the risk of betrayal (Baier 1986, p. 235). I will argue that willingly making oneself vulnerable towards someone implies placing moral responsibility on that someone. Furthermore, I will argue that from the trustee's side, committing to not abuse others' vulnerability entails assuming moral responsibility. The trustor's vulnerability creates an expectation that the trustee will act in alignment with the trustor's motives and interests. This expectation is not only a practical one but also an ethical one, as the trustor's act of placing trust signals a reliance on the trustee's integrity and intentions. Trust, therefore, creates an implicit moral contract where the trustee is expected to recognize and honor the trustor's vulnerability. Because of this, responsibility attribution becomes normatively necessary in contexts where people's lives can be deeply affected, as is often the case with the use of AI. AI systems are in use in contexts in which the affected parties are highly vulnerable to the outcomes of the AI system, or to decisions derived from such outcomes. Some examples of such high-stakes contexts include healthcare or personal data management. Researching the moral aspect of trust in AI is not only meant to contribute to upholding ethical standards but also to foster a sustainable and just integration of AI into the fabric of society. Because of how fitting the moral implications of trust are to the contexts in which artificial trustees are already in use, I will argue that trust is a suitable goal to aim for when shaping human-AI relationships.

---

<sup>1</sup>I follow the use of 'moral significance' as employed by Persson (2021). In his words, "[s]omething has moral significance just in case it morally counts for its own sake, or is something that must be taken into consideration in itself when moral judgments about what ought or ought not to be done are made" (p. 108). In this sense, some relationships, such as trust, are morally significant because the relationship itself becomes relevant when making moral judgments about the actions of those involved, by virtue of that relationship.

To understand the responsibility that comes with trust, I aim to analyse the roles of both trustor and trustee. In trusting relationships, the trustor benefits from the relationship whilst, despite being in a vulnerable position, their expectations should be satisfied by the trustee. But what does the trustee get out of such a relationship? Even though being trusted demands responsibility, generally, people aim to be perceived as trustworthy. Both in ordinary language and in part of the trust literature, trustworthiness is considered a virtue. People who are perceived as trustworthy are socially valued and their contributions to diverse endeavours tend to be demanded. In consequence, trustworthiness brings a variety of social goods to those who exhibit it. Being perceived as trustworthy can grant one a job, a friendship or a good reputation. Thus, the (seemly) trustworthy trustee plays an active role in society that provides them with both social goods and personal satisfaction.

As in interpersonal cases, trust in AI involves the trustor's expectations and the accompanying risk of those expectations not being met. The human trustor expects that the system will provide accurate results; in other words, they expect certain competence. In the case of artificial trustees, seeming trustworthy does not directly grant social benefits to them; but it does to the people behind them. In chapter 2, I described AI as a socio-technical tool (STT), meaning that AI systems consist not only of software and hardware but of social components such as the developers and model creators. I will argue that human trustees are morally responsible for how they present themselves to trustors and behave accordingly. Similarly, artificial intelligent systems should also fulfil the expectations they generate. Failing to do so would be morally reprehensible. Hence the need to trace back responsibility to the humans who have built the systems and those who have presented them as trustworthy to the users and society.

The chapter is structured as follows. In section 6.2, I will lay down the foundation for my argument by introducing my view of moral responsibility as a relational concept. Instead of focusing on the consequences of actions, I will concentrate on the relationships between the parties involved to ascribe responsibility. This approach is particularly suitable when one party is an AI. I will explore the normative expectations that accompany trusting relationships and revisit the idea of trust as an invitation, first introduced in chapter 2.<sup>2</sup> In section 6.3, I will explore the moral component of trust, specifically in the context of AI. Building on the concept of trust as an invitation, I will explain the role of the trustee in the trustor's belief-formation process. The core idea is that the trustee has the moral responsibility to present an accurate image of their trustworthiness to the trustor. I will clarify what constitutes such an image and why its accurate portrayal is morally significant, particularly for artificial trustees. In section 6.4, I will revisit the conceptualisation of AI as a STT presented in chapter 2.<sup>3</sup> By this, I mean to dissect the different parts of AI trustees, both social and technical, and clarify to which and how moral responsibility can be attributed. Finally, in section 6.5, I will revisit my account of trust as a belief-based disposition (BBD) developed in chapter 3. I will explore how the moral component of trust works in that framework and its relevance in the concept of appropriate

---

<sup>2</sup>Section 2.5.3

<sup>3</sup>Section 2.3

trust.<sup>4</sup> My aim is to connect the view of moral responsibility in human-AI trusting relationships with the account of trust developed across previous chapters. By this, I attempt to provide a conceptual framework that helps to understand trust in AI, together with its moral implications.

---

## 6.2 Trusting relationships and moral responsibility

In this section, I will argue that trusting relationships imply moral responsibility. To do so, I will start by introducing my view of moral responsibility as a relational concept. This view contrasts with the understanding of moral responsibility as a state linked solely to specific actions, independent of the relationship between the actor and the recipient of such actions. While both perspectives are complementary, it is crucial to recognize that moral responsibility inherently involves the relational dynamics between individuals. I will then apply this framework to the specific case of trusting relationships, focusing on the normative expectations that accompany this kind of relationship and on the idea that trust constitutes an invitation from the trustee to the trustor to delegate some role or action to the trustee.

### 6.2.1 Moral responsibility as a relational concept

Moral responsibility is a common concept both in ethics and daily life that refers to the status of being accountable for one's actions in light of moral standards (Talbert 2024). Moral responsibility can be understood in two ways:

- a) Moral responsibility emerges from *actions*. This perspective focuses on actions' consequences, and it aligns with the premise that moral agents are object of praise or blame for the consequences of their actions (Sartorio 2016, p. 77), regardless of their relationship with the recipient of such actions.
- b) Moral responsibility emerges from *relationships*. This perspective sees moral responsibility as arising from interpersonal relationships and social contexts (Strawson 1962, p. 116; Watson 1996, pp. 266–67; McKenna 2012, pp. 7–9, 78).<sup>5</sup> It emphasizes the importance of relational dynamics, mutual expectations, and social roles. In this view, individuals are accountable not only for their actions but also for their relational commitments and the impact they have on others.

These two perspectives are not mutually exclusive. Often, moral responsibility involves a combination of both actions and relationships. For example, someone might be held accountable for a specific action (such as breaking a promise) because

---

<sup>4</sup>See definition 3.6 in chapter 3.

<sup>5</sup>Authors such as Watson (1996) distinguish between *attributability* and *accountability*. *Attributability* pertains to assessing an agent's moral worth based on their actions, which would reveal aspects of their moral character. By contrast, *accountability* relates not only to the agent's actions, but to their management of others' expectations. In this sense, Watson's distinction fits the perspectives on moral responsibility described above.

it violates the expectations inherent in a relationship. Understanding moral responsibility in both ways provides a more comprehensive view of ethical behaviour and accountability in human interactions. This is the compatibility view I will adopt, integrating both actions and relationships as central to moral responsibility. In light of the characterisation of trust I have provided so far (see chapter 3), this chapter will focus on the relational aspect of moral responsibility. This means that certain relationships imply accountability from one party to another, not just based on specific actions, but by virtue of the relationship itself.

My view on moral responsibility is influenced by Strawson's (1962)'s "Freedom and Resentment".<sup>6</sup> Strawson addresses moral responsibility from a participant standpoint, according to which we regard others as appropriate subjects of *reactive attitudes*. Reactive attitudes are emotional responses that people naturally have toward the actions and intentions of others in interpersonal relationships. These attitudes include feelings like resentment, gratitude, forgiveness, and moral indignation. An agent does not hold a reactive attitude merely as a consequence of actions that affect them, but in the context of the relationship such an agent has with the performer of the action. For example, I feel resentful towards someone who harms me only if I cared about that person and thought they cared about me, but nevertheless, they intentionally harmed me. On the contrary, if the harm is accidental, resentment would not be an appropriate reaction. Strawson argues that these reactions are central to our practices of holding each other morally responsible. They are not merely judgments about behaviour but are deeply tied to the intentions and motives that the actions reflect. Thus, reactive attitudes ground moral responsibility in people's natural emotional responses and in the context of interpersonal relationships.

In sum, I understand moral responsibility as being accountable for one's actions within the framework of the relationship between the agent who acts and the recipient of those actions. This framework encompasses the implicit and explicit commitments between the agent and the recipient, along with the expectations arising from these commitments. For example, in a doctor-patient relationship, the doctor is morally responsible for the patient. This responsibility arises not only from specific actions, such as providing a diagnosis or conducting tests, but also from the inherent moral obligation that partly defines this kind of relationship. This moral responsibility exists by virtue of the relational context itself, emphasizing the significance of the commitments and expectations that characterize such interactions.

### 6.2.2 Trust as normative expectation

In chapter 3, I defined trust as the belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  to a trustee  $B$  (see definition 3.4). That is, trusting someone means to expect a certain behaviour from them. The trustor expects the trustee to successfully perform the task or role that could<sup>7</sup> be delegated to them. Whether this is a normative expectation is a topic of debate in the literature on trust. While the

<sup>6</sup>I will follow the page numeration from Hieronymi's (2020) reprint in *Freedom, Resentment, and the Metaphysics of Morals* (pp. 107-134).

<sup>7</sup>I write 'could be delegated' and not 'is delegated' because trust deals with hypothetical scenarios, not necessarily with actual delegation. For further details on this idea, see section 1.3, in chapter 1.

debate generally leans towards the view that trust involves a normative expectation (Baier [1986]; Jones [1996]; Cogley [2012]; Hawley [2014]), there is still some disagreement among authors (Gambetta [1988]; Hardin [2002]). Authors who disagree normally argue that trust is not about expecting someone to do what they should. Instead, these authors defend that, to establish a trusting relationship, there just needs to be a goal alignment between trustor and trustee, regardless of the kind of goal. Thus, trust is seen as a pre-condition for cooperation (Gambetta [1988], p. 217). A classic example is the society of robbers and murders described by Adam Smith ([1759], p. 86)<sup>8</sup>, in which even those who act against moral norms must trust, at least, each other.

In my view, trust inherently involves a normative dimension, which includes an aspect of ‘internal normativity’ specific to the relationship between the trustor and trustee. By internal normativity, I refer to expectations specific to the trusting relationship, where the trustor expects the trustee to adhere to a shared norm. In other words, the trustor has *normative expectations* towards the trustee.<sup>9</sup> There is an internal norm that ought to be respected by the trustee. When this shared norm is broken, the trustor can feel betrayed.

With this kind of internal normativity (not necessarily accepted by a wider social group) come several problems. The most pressing one is how explicit the trustor-trustee’s shared norms are. Ideally, every trustor’s expectations would be made explicit to the trustee so both parties are clear about what each expects from the other and how likely it is that these expectations will be met. However, such cases describe contracts rather than trusting relationships. As described in the first two chapters of this thesis, trust encompasses a more ambiguous realm that allows for risk, uncertainty, and betrayal. Closely related to this problem is the issue of the trustee’s awareness of the trustor’s expectations. Since trust relationships are based on internal norms that may not be shared by a wider group of people, if those norms are not made explicit by the trustor, they may not be obvious to the trustee. Then, misunderstandings and accidental betrayals can occur, stemming from the trustee’s ignorance rather than ill intention or actual motive misalignment. Finally, a third potential problem that can arise from the internal normativity that characterises trust is unreasonable expectations. Sometimes, even when the trustee is aware of the trustor’s expectations, the trustee may deem such expectations unreasonable. Since trust encompasses a very restricted internal normativity, it does not really matter whether the trustor’s expectations are actually unreasonable or not, but whether the trustee considers them to be so. For example, let us go back to the neighbours Claire and Christian from case study C.<sup>10</sup> Imagine a scenario in which Claire trusts Christian to take care of her cat and she expects Christian not only to care for the animal and feed it regularly, but also that Christian takes the cat to the vet and that he pays for the cat’s medication out of pocket since the animal

---

<sup>8</sup>In that passage, Adam Smith was originally describing pre-conditions for a society (of any kind) to exist. To put the citation above in context, here is the literal quote: “Society, however, cannot subsist among those who are at all times ready to hurt and injure one another. [...] If there is any society among robbers and murderers, they must at least, according to the trite observation, abstain from robbing and murdering one another” (Smith [1759], p. 86).

<sup>9</sup>For a different interpretation of the normative dimension of trust, see my critique on Ryan’s view on normative motives in section [5.3] chapter [5].

<sup>10</sup>See section [2.4.3] in chapter [2].

has ongoing health issues that Claire is aware of. Christian is informed of this situation because Claire has verbalised it, just before leaving. Even with Christian's knowledge, Claire demands are considered unreasonable. Claire's trust has become a burden to Christian since she made him participate in a norm that Christian never agreed with. In this kind of situation, trust can be used as a tool for manipulation.

However, not everything is problematic about the normative expectations implicit in trusting relationships. Having a more restricted sense of normativity shared between the trustor and trustee can enhance more widely accepted social norms. The internal normativity between trustor and trustee does not necessarily contradict general norms or lead to manipulation. The shared norms can align with social norms, and additional expectations from a trustor can reinforce compliance with these broader norms. For example, when I trust ChatGPT to correct my English, I expect it to generate grammatically correct sentences and identify mistakes in my texts. This expectation aligns with the broader norm of valuing clear and correct communication. By meeting my expectations, ChatGPT not only fulfills the specific normativity of our trust relationship but also upholds the wider social norm of linguistic accuracy. This alignment shows that internal normativity within a trust relationship can support and enhance general social expectations. This way, meeting the trustor's expectations motivates the trustee to conform not only to the shared norm that they have with the trustor but also to broader social norms, such as moral ones.

### 6.2.3 Trust as an invitation

In chapter 2, I provided a brief overview of different accounts of interpersonal trust. In section 2.5.3, I covered views that conceptualised trust as a contract. According to these views, trust can be understood as a non-enforceable implicit contract in which the trustee commits themselves to fulfil the trustor's expectations. From these views, I took the idea of seeing trust as an invitation. That is, a trustee can invite a potential trustor to trust them with their behaviour, or explicitly ask them to do so. This applies both to interpersonal and institutional trust since both individuals and institutions normally seek to be seen as trustworthy.

A bit more ought to be said about inviting others to trust. When it comes to explicitly inviting with words potential trustors to trust, the trustee puts themselves in a paradoxical situation. Since trust is non-voluntary,<sup>11</sup> asking others to trust us seems fruitless: either the trustor already trusts the trustee, in which case their invitation serves only as reassurance, or they do not, in which case they may just reply with the daunting "I wish I could trust you" (Baier 1986, p. 244). But trustees can also invite potential trustors to trust them not with their words, but implicitly with their behaviour.

As I developed in chapter 3, the trustor trusts the trustee when the former believes that the latter is trustworthy. This trust belief can be based on a variety

---

<sup>11</sup>Most authors consider trust as non-voluntary, in a similar way as belief (Baier 1986; Mills 1998; Hieronymi 2008; Keren 2014). However, a noteworthy exception can be found in Holton (1994), who characterizes trust as not requiring the trustor's belief in the trustee's trustworthiness. Since my account on trust is founded on an entirely different premise, I will not delve further into voluntary approaches to trust.

of reasons, often derived from observing the trustee's behaviour. Thus, the trustor believes the trustee is trustworthy because the trustee presents themselves as such. Through such a presentation, the trustee invites the trustor to trust. In the case of Claire and Christian, Christian invites Claire to trust him by presenting himself as a trustworthy person, at least when it comes to caring for cats. He invites Claire to trust him both with his words –by telling her about his own cats and how well taken care of they are– and with his behaviour –by being seen taking his cats to the vet regularly, for example–. Christian is responsible for the image of trustworthiness he projects, especially when it is deliberately aimed at Claire. If he often talked about how well he cares for his cats while actually being cruel to them, he would be portraying himself in a misleading way, which is morally reprehensible. Misleading others about one's trustworthiness is morally wrong because it causes them to act based on false information, leading to decisions they would not make if they knew the true facts. Based on this argument, I consider trust to be a moral concept since, in trusting relationships, the trustee has a moral responsibility towards the trustor. This responsibility will be weaker or stronger depending on the specifics of the particular trust relationship. But, at the very least, the trustee owes the trustor not to mislead them into believing that the trustee will behave in a way that most likely they will not. In other words, the trustee has the moral responsibility of not making empty or misleading invitations to the trustor that may lead to the trustor believing that they should delegate some task to the trustee when the trustee is either not capable or willing (or both) to successfully perform such a task.

---

## 6.3 Trust in AI as a moral concept

So far, I have argued that in trusting relationships, the trustee has the moral responsibility to present an accurate image of their own trustworthiness to the trustor. In this section, I will explain what constitutes such an image and why its accurate portrayal is morally significant in the particular case of artificial trustees.

### 6.3.1 Portraying trustworthiness

As I argued in the last section, the trustee can present themselves as trustworthy through their testimony or their behaviour. In the case of AI systems, typically the human developers are the ones who present the system as trustworthy.<sup>12</sup> Since I consider the developers as part of the social component of AI, I won't be making further distinctions between the developers and the technical parts of the system, but rather consider that in trusting human-AI relationships the trustee is AI as a STT.<sup>13</sup>

---

<sup>12</sup>It exists the possibility that an AI system presents 'itself' as trustworthy by communicating its own capabilities or consistently providing accurate results. However, even in those cases, such communications and accurate results are the indirect product of the developers' design of the system. To simplify my argument, I will stick to the narrative that the (technical component of the) AI system exhibits behaviours and the human developers (as the social component of the AI system) portray such behaviours as (un)trustworthy.

<sup>13</sup>For details, see section 2.3 in chapter 2

Similarly to the way in which human trustees can present themselves as trustworthy to potential trustors, AI developers<sup>14</sup> present their systems in a certain way. Let us revisit the example of ChatGPT as a trustee as exposed at the end of section 3.3.2 in chapter 3. In 2023, OpenAI promoted ChatGPT-4 as a language model able to exhibit “human-level performance on various professional and academic benchmarks” (OpenAI 2023, p. 1). With statements like this, they created certain expectations in users and lead them to certain beliefs, such as the belief that ChatGPT is more likely to produce a grammatically correct sentence in English than myself. As was the case for interpersonal trusting relationships, OpenAI bears some responsibility for shaping users’ perception of ChatGPT.<sup>15</sup> If users ultimately place their trust in the system, OpenAI shares part of the responsibility for fostering that trust and the users’ willingness to be vulnerable. In other words, OpenAI is responsible for the image of trustworthiness that they portrayed of ChatGPT.

A crucial question is what an image of trustworthiness consists of. As developed in chapter 3,<sup>16</sup> a trustee is trustworthy for a trustor when it is reasonable for the trustor to delegate some task or role to the trustee. Doing so is reasonable if the trustor has access to rational reasons that indicate that the trustee would most likely perform the delegated task successfully, moved by motives deemed appropriate by the trustor. From this definition, it can be concluded that portraying trustworthiness consists of offering reasons to believe that the trustee has the capability to do something and certain motives to do it. The question is: how can AI systems –or more accurately, their developers– offer such reasons? In the case of ChatGPT, OpenAI clearly stated what the system is capable of, leading users to expect such skills. This is an example of portraying an image of trustworthiness through testimony. To understand how human users can access artificial trustees’ motives, let me expand on the view of motives that I offered towards the end of chapter 2 (section 2.7).

As previously pointed out, a trustee is considered trustworthy not only if they are capable of successfully performing the task to be delegated, but also if there is a *perceived motive alignment* between the trustor and the trustee. In a trusting relationship, the trustor believes that their motives align with the trustee’s. If the trustor’s perception of the trustee matches reality, most likely trust will be warranted (as long as other requirements such as the trustee’s competence are also satisfied). However, the trustor’s perception of the trustee’s motives may be distorted, leading the trustor to place unwarranted trust in the trustee. For that reason, I specified that in trusting relationships there is a *perceived* alignment of motives, not necessarily a real alignment.<sup>17</sup>

When it comes to AI systems, their ‘motives’ to act are the processes they follow

---

<sup>14</sup>I will focus on developers as their responsibility for the systems they create is generally accepted. Listing all potentially responsible parties for AI’s actions goes beyond the scope of this thesis. Nonetheless, policymakers, stakeholders, etc. could also bear responsibility, to varying degrees, depending on the circumstances.

<sup>15</sup>For an analysis of Twitter users’ perception of ChatGPT during one of its peak of popularity (from November 2022 to March 2023), see Heaton et al. (2024).

<sup>16</sup>See definition 3.2.

<sup>17</sup>From now on, I will assume this clarification and not insist on it in virtue of clearer readability.

to achieve results.<sup>18</sup> That is, their trustworthiness depends not only on whether they are capable to minimise a loss function but also on which process they follow to do so. Depending on whether the motives that a human trustor has to delegate a task to an AI system align with the process followed by the system (its ‘motives’), such process would be considered –by the trustor– appropriate or not. I use ‘appropriate processes’ in a loose sense, meaning that the loss function was not minimised by chance or due to fortunate correlation. Rather, the system provides accurate results because it ‘pursues’ the objectives that its human trustors want it to pursue. Remember case A from section 2.4.1 (chapter 2). The diagnostic system is worthy of trust if it successfully detects tumours by focusing on the areas of the MRI where the tumours actually are. By contrast, if the NN predicted where the tumours are by focusing on some other non-causal correlation factor, the system should not be trusted. In other words, if the human trustor approves the process (focusing on the pixels corresponding to tumours) behind the NN’s predictions, the system oughts to be trusted.

When evaluating whether a trustee is worthy of trust for a specific task, the existence of motive alignment varies from trustor to trustor since different trustors deem different motives appropriate. For example, let us consider different trustors presented with an automated recruiting system capable of finding the best candidate for a specific job vacancy.<sup>19</sup> All trustors agree that the system is capable of providing accurate results, as it has consistently chosen the best candidates over time: that is, the system has portrayed its trustworthiness through its ‘behaviour’ (since it is an artificial trustee, through its performance). However, each trustor considers appropriate different motives to guide the process that produce such accurate results. Trustor  $A'$  prioritizes avoiding gender bias and would only trust the system if it has been trained on sufficiently diverse data to prevent biased decision-making. By contrast, trustor  $A''$  does not prioritize gender bias and prefers faster results, valuing efficiency and speed over the diversity of the dataset. Trustor  $A''$  will trust the system if its developers present it as being driven by values such as efficiency and speed. Meanwhile, trustor  $A'$  will consider whether the system’s developers have promoted it as not only accurate but also non-biased. This example illustrates that trustworthiness depends on both the trustee’s perceived capabilities and whether the trustee is portrayed in such a way that leads the trustor to believe that there is an alignment of their motives with those of the trustor. Different trustors will have different criteria for this alignment, which mirrors the subjective nature of trust.

Misleading potential trustors about either the trustee’s capabilities or their motives to act is morally reprehensible. In the next section, I go into the specifics of why this is the case when the trustee is an AI.

### 6.3.2 The moral significance of portraying of trustworthiness

As anticipated in section 6.2.3, offering a false self-image of trustworthiness is morally wrong because it leads trustors to act based on false information. Specifically, it

---

<sup>18</sup>For further details on my use of the term ‘motive’, see section 2.7 in chapter 2 or the corresponding entry in the Glossary for a concise definition.

<sup>19</sup>For additional details on this example, see section 2.4 in chapter 2

causes them to delegate tasks to trustees when they would not have done so if they had a more accurate perception of reality. This is morally wrong since it constitutes a form of manipulation. Manipulation, in this context, is understood as influencing a target using methods without considering whether those methods provide the person with good reasons to act, think, or feel as the manipulator desires (Klenk 2022, p. 97).

In human-AI trusting relationships, building a faithful image of the artificial trustee is particularly pressing, given the high-stakes contexts in which AI is already in use. A high-stakes scenario is a complex situation characterized by ambiguous factors, such as an emergency event, where decision-makers must make critical decisions under pressure. These scenarios have a low probability of occurrence but carry severe consequences, such as loss of life, if not handled correctly. In such contexts, decision-makers need practical strategies and comprehensive information to guide their actions and prevent catastrophic outcomes (Sahoh and Choksuriwong 2023, pp. 7827–8). In this kind of decision-making, it is crucial being able to find relevant information among a large and complex amount of data. Because of that, AI systems are being used to assist some of these decisions, such as medical diagnosis (e.g. IBM Watson for Oncology), loan grants (e.g. ZestFinance), sentence attribution (e.g. COMPAS), etc. High-stakes decision-making emphasizes human safety and security; therefore, it is necessary to attribute (some degree of) responsibility for the outcomes.

The need for responsibility attribution in these contexts arises from the increased vulnerability that trustors experience. In the case of artificial trustees, the trustor is placed in a position of vulnerability due to the lack of control over an outcome that will deeply affect them. Furthermore, the trustor often lacks understanding of the process the artificial trustee follows to reach a result, adding an extra layer of inaccessibility to the decision-making, of which the trustor is the subject. As it was pointed out in chapter 4, opacity can arise from technical illiteracy, where certain systems are inaccessible to individuals without the necessary technical background (Burrell 2016, p. 4). Being subject to an opaque decision-making process is an added factor that further aggravates the trustor's vulnerability. For example, imagine a hospital using an AI system to decide whether a patient qualifies for an urgent medical treatment, such as prioritizing patients for organ transplants or emergency surgeries. The patient, as the trustor, has no control over the AI's decision-making process and may not fully understand how the AI weighs factors such as medical history, urgency, or other health data. The lack of transparency and control over these decisions places the patient in an extremely vulnerable position, as the AI's recommendation could directly affect their chances of survival or recovery.<sup>20</sup> Given the increased vulnerability that results from high-stakes scenarios, when AI is used in these contexts, it is required to use systems worthy of trust.

I emphasize the context of use as a condition for requiring trust in some AI-human relationships, as this is not always the case. AI systems are utilized in various domains, and often mere reliance is sufficient. Take as an example AI-based predictive text, often present in smartphones' keyboards or email services. Users can rely on this function without necessarily trusting it. If the predictive text fails and I send

---

<sup>20</sup>For further details on how some uses of AI systems can exploit users' vulnerability in adversarial contexts, see Bordt et al. (2022, p. 892).

an email signing with “kind *regrets*” instead of “kind *regards*”, I will not feel betrayed or manipulated by the system. I merely used a tool and it failed, similar to when I use a ladder and it collapses under my weight. However, when individuals affected by AI’s outcomes are particularly vulnerable, responsibility attribution becomes essential, necessitating a trusting relationship. The moral component of trust makes it a suitable goal in such contexts. In sum, it is normatively necessary to aim for trust when shaping human-AI relationships, especially in high-stakes scenarios, because these situations require a degree of responsibility attribution (HLEG 2019, p. 22). The moral component of trust aligns with this need.

Responsibility attribution is a complex task that does not only involve the system’s technical success but the process behind that success (or failure). Trusting AI does not only mean relying on its potential to minimize a loss function. It means to believe that the loss function is likely to be minimised due to an appropriate process. As I argued in the previous section, this means that a system will be considered trustworthy not only in virtue of its capabilities, but also attending to its working processes or ‘motives’. A rich debate exists on who is responsible for AI systems’ performance, with no clear consensus due to the difficulty to trace back the system’s output to human agents (Matthias 2003; Sparrow 2012; Coeckelbergh 2020; Kiener 2022). There are indeed human developers shaping these performances and ensuring that the systems’ goals align with the human users’ goals. One could rightly argue, and this is often done (Mittelstadt et al. 2016; Dignum et al. 2018), that part of the responsibility for the system’s actions can be traced back to them. However, following the previous discussion, this kind of responsibility attribution is not only due to the causal connection between the developers’ actions and the systems’ outcomes but also due to the expectations that the developers generate regarding those outcomes. More generally, understanding AI as complex socio-technical systems, it is often unclear who causes what. In other words, which human components of the systems are the (partial) cause of specific system’s actions. The attempt to trace human responsibility back for the system’s actions is a complex endeavour. For this reason, I propose instead to focus on who is responsible for the trustor’s expectations and their effect on the trustor’s use of the system.<sup>21</sup> Human trustors use AI according to the skills that they expect the system to have. When developing a system, its designers (among others) portray a system’s image of the system’s capabilities and potential. They are responsible for such an image and have the moral responsibility to promote the system in a way that matches reality. In sum, if a tumour detection system is promoted as able to detect tumours, its developers have the responsibility for the system to do precisely that. A system that happens to find tumours *by chance* should not be promoted as trustworthy and, consequently, should not be trusted.

---

<sup>21</sup>Note that, sometimes, the responsible for the trustor’s expectations may be solely the trustor themselves due to some cognitive bias (Kirby, Kirton, and Crean 2018, p. 78). However, this possibility does not exempt developers and stakeholders from portraying an accurate image of the systems’ capabilities.

---

## 6.4 Artificial trustees and responsibility

As previously pointed out, artificial trustees have both a technical and a social component. The technical component comprises the AI system itself, understood as a set of software and hardware. The social component consists of the human agents responsible for the software and hardware, whether by designing, developing, training, marketing, approving, or distributing it. In human-AI trusting relationships, the trustee is the AI, understood as a STT that includes both the social and technical parts. However, the responsibility attribution that accompanies trust can only be applied to the social part of the system, i.e., the humans behind it (Simon [2010](#), p. 61).

At the moment, working in AI can be very lucrative. Developing AI that is considered trustworthy can lead to a profitable business, as well as professional and social recognition. These are legitimate social benefits, as long as the systems are as trustworthy as they are portrayed to be. Human agents are responsible for such an image of trustworthiness, and as the social part of the AI system, they also constitute a part of the artificial trustee.

In this chapter, I have offered a view on moral responsibility focused on the relationship between the parties, rather than the consequences of their actions. This view suits human-AI relationships because, often, there is not an agent responsible for the consequences of a particular action. Consider again the use of AI in medical diagnostics. If an AI system makes an incorrect prediction, responsibility cannot be attributed to the technical part of the system, i.e., the AI itself. AI systems lack agency, and therefore, no traditional sense of responsibility can be attributed to them<sup>[22](#)</sup>. Sometimes, responsibility for the consequences of AI's actions can be traced back to its social components, that is, the humans who indirectly caused such actions. However, this is a very complicated task. In the case of a diagnostic system such as the tumour detection system from case A,<sup>[23](#)</sup> who is responsible for the NN focusing on pixels that did not correspond to an actual tumour? It seems unfair to blame the designers who built the software and tested it without detecting errors during the training phase. Even if they bear *a part* of the responsibility, it is very complicated to judge how to distribute responsibility among specific individuals and to determine the weight of each person's actions that eventually contributed to a wrong diagnosis (that may have happened years after those developers worked on the particular system). As Matthias ([2003](#)) pointed out over 20 years ago, there is a responsibility gap when trying to ascribe responsibility for the actions of non-human quasi-agents. It is not my aim to bridge that gap in this thesis. Instead, I have focused on who is responsible for the trustor's expectations and how these expectations influence the trustor's use of the system. In the case of AI, some of the humans belonging to

---

<sup>22</sup>Note that, as I clarified in section [2.3.2](#) (chapter [2](#)), I consider AI systems to be quasi-agents, not full agents. In consequence, I agree with Nickel, Franssen, and Kroes ([2010](#)) Ryan2020 regarding AI's lack of agency, even though I criticised their final conclusions. The core of my disagreement lays of the way of understanding AI; while for them AI systems are just its technical component, I conceptualise it as a STT, and it is this tool as a whole what I take to be the artificial trustee.

<sup>23</sup>See section [2.4.1](#) (chapter [2](#))

the social part of the system are responsible for the image of trustworthiness that the users have of the system. They have the moral responsibility of portraying an accurate image of the system's capabilities and motivations.

A fair inquiry about the argument above is, if only the social part of AI systems can be responsible for the system's outcomes, how is trust in AI different from institutional trust? In the latter case, individuals perform actions abiding by an institution that bears responsibility only in an abstract sense. In a more practical sense, the members of the institution are the ones who will be judged for their actions, even if they are not judged in the same way they would be if they were personal decisions rather than institutional roles. Similarly, in the context of AI, the individuals behind the AI system (developers, engineers, data scientists, etc.) are the ones who can be judged for the system's performance. However, trust in AI differs from institutional trust because the actions taken by AI systems often involve complex algorithms and automated processes that lack direct human oversight at every step. This is the reason for talking about quasi-agency. This introduces a layer of technical opacity that does not exist in traditional institutional roles, where human decision-making is more transparent and accountable. Moreover, AI systems operate based on data-driven models that continuously evolve, making it challenging to pinpoint responsibility at any single moment. The dynamic nature of AI systems is influenced by two main factors: continuous learning and the rapid pace at which models evolve. As a result, trust in AI must constantly adapt to these changing circumstances and capabilities. This stands in contrast to institutional trust, which operates within a more stable and predictable framework. Nevertheless, establishing the parallel between trust in AI and institutional trust is helpful in understanding the former. In particular, this parallel helps to overcome the comparison between trust in AI and interpersonal trust, whose features cannot always be compared one on one.<sup>24</sup>

The parallel with institutional trust is also pertinent since, often, the people responsible for the technical part of AI systems belong to companies or organisations. In this case, trust in AI and interpersonal trust intertwine, making it even more complicated to properly identify the trustee in the trusting relationships that are established with the users. At this point, it may be helpful to remember the characterisation of AI as a STT that I developed in chapter 2, section 2.3.

Traditional tools, such as a hammer, are artefacts that can be used to perform a specific task, such as driving nails. Normally, tools are used in a passive way. The user of the tool is considered the performer of whatever task the tool is being used for, while the tool is merely a means: I drive a nail using a hammer, but the hammer itself does not really do anything. In contrast with this low-level notion of a tool, there are more complex artefacts that belong to socio-technical systems. Take a coffee machine as an example. Like the hammer, the coffee machine needs a user; it cannot operate itself. However, the responsibility for the machine's actions could extend beyond the user, especially in cases of accidents or manufacturing defects. A coffee machine has no autonomy to make coffee, but it is nevertheless integrated into a network of human and technical elements that collectively influence its use,

---

<sup>24</sup>For details about this, see sections 2.4 and 2.5, in chapter 2

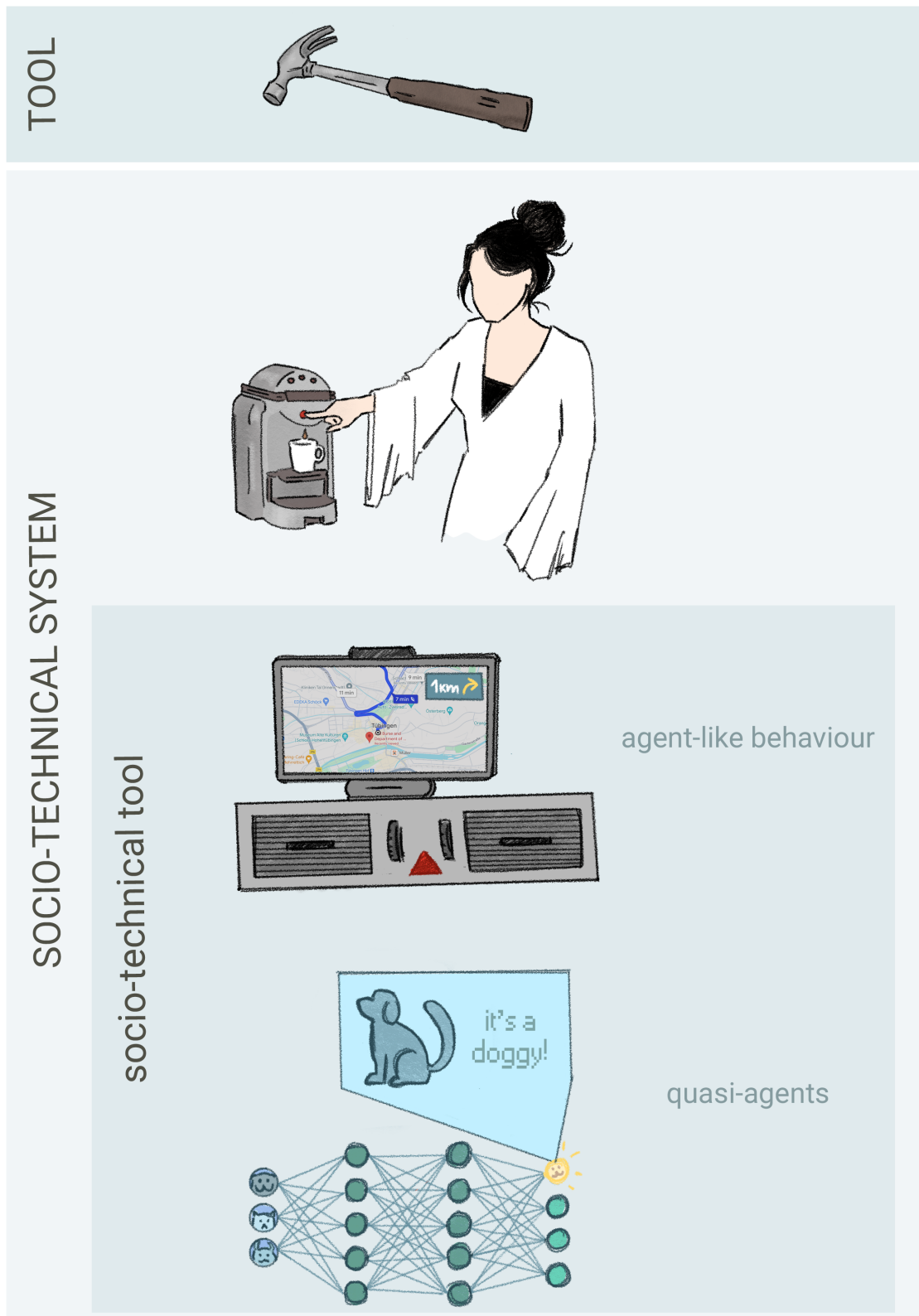


Figure 6.1: In this thesis, I have used the term *socio-technical tool* to refer to the kind of technology that needs of a social component to work, besides being part of a wider socio-technical system (for clarifications regarding this terminology, see figure 2.3 in chapter 2).

maintenance, and impact on society. For example, in the case of capsule machines, the user remains dependent on the manufacturer to continue purchasing capsules for the machine. Depending on user feedback and market dynamics, the manufacturer would offer a certain range of flavours and options. Therefore, the task of making coffee goes beyond the user and the tool, involving a broader network of both social and technical elements.

If we go a step further, we encounter autonomous machines, which I have labelled as STTs and are also ascribed to socio-technical systems (see figure ??). The term STT does not exclusively refer to AI. As Nickel, Franssen, and Kroes (2010, p. 439) described, there are other artefacts that exhibit agent-like behaviour, such as navigation systems.<sup>25</sup> The difference between this kind of technology and what I consider to be quasi-agents is that in the latter case, there is an increased layer of technical opacity: the outcomes of some AI systems actually escape the control of their designers and cannot be predicted by them. Both kinds of performances mimic agency, but in the case of quasi-agents, the imitation can reach actual autonomy and eventually lead users to interact with the machine ‘as if’ they were interacting with a full agent.<sup>26</sup>

AI is not the first case of technology to which the users relate in such a way that the manufacturer bears responsibility for how the product is portrayed. However, the case of AI is novel in the sense that, even if responsibility can only be ascribed to humans, the user *trusts* the tool as a whole. This is so because their interactions with the technical parts of the system mimic human behaviour in an unprecedented way, making users relate to the technology in different ways that they do with other tools. Unlike other technologies, it is possible to partially trust AI systems because we genuinely delegate tasks to them. When using other artefacts, such as a car, our skills are merely extended, and we remain the ultimate performers, guiding and controlling the action. By contrast, when using AI, we recognise another quasi-agency; we let go and truly delegate tasks to the AI system. This delegation is a significant departure from the use of traditional tools, where human intervention is continuously required.

This concept aligns with Kirton’s (2020) notion of trust as attachment. Trust is a social response to our intrinsic need to form attachments and feel supported. According to Kirton, we need to trust others to avoid feeling overwhelmed by life; we need to feel that someone has our back. Without such trust, the constant vigilance required would lead to anxiety and exhaustion. This fundamental need to ‘let go’ and trust can be fulfilled by AI systems. When we trust AI, we delegate tasks to these systems, easing our cognitive load and allowing us to focus on other activities. This delegation mirrors the trust we place in human agents, tapping into the same psychological mechanisms that drive us to seek and maintain social bonds.

---

<sup>25</sup>I am referring to simple GPS devices that rely on static map data and satellite positioning to navigate, such as the ones that Nickel, Franssen, and Kroes (2010) seem to have in mind.

<sup>26</sup>See section 2.3 in chapter 2.

## 6.5 Revisiting the BBD account

In chapter 3, I developed my account on trust as a belief-based disposition (BBD). I argued that there is a specific kind of trust that should be placed in artificial trustees, which I labelled as *appropriate trust*. The main difference between trust in a more general sense and appropriate trust is that, in the latter, the trustor's trust belief must be justified. Let us briefly recall the definition that I proposed in chapter 3:

**Definition 3.6** *Appropriate trust* is the justified belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold

- (i')  $A$  has a continuing *justified* belief that  $B$  is trustworthy.
- (ii') If  $A$  wished or needed  $B$  to do  $\delta$ , then  $A$  would be willing to delegate  $\delta$  to  $B$ .

In sum, the trustor appropriately trusts the trustee to do some task  $\delta$  when the trustor has access to rational reasons that support the conclusion that the trustee is likely to successfully perform  $\delta$ . Those reasons should justify the trustor's beliefs about the trustee being capable of doing  $\delta$  and motivated to do so by motives deemed appropriate by the trustor. When the trustor holds this kind of justified trust belief, the trustor is disposed to delegate  $\delta$  to the trustee. This means that, in case that the trustor wished or wanted to delegate  $\delta$  to the trustee, the trustor would be willing to do so –since it would be coherent with their belief system.<sup>27</sup>

Previously, I argued that appropriate trust is *normatively desirable* because it increases the likelihood of placing trust in trustworthy trustees. This should be the goal for all types of trustees. However, as discussed earlier in this chapter, the high stakes and increased vulnerability often accompanying human-AI trusting relationships make the need for justified trust beliefs even more pressing.

In this chapter, I have delved into the concept of trust as a normative expectation. Initially, when I characterised trust as a belief-based disposition, I did not explore its moral component. By now, this component should be clearer. Belief-based dispositions are held by an agent  $A$  about another agent  $B$  –in the case of trust, the trustor and trustee, respectively. In a trusting relationship, the trustor holds beliefs about the trustee that prompt the trustor to act in a certain way. We have seen that the trustee has a moral responsibility towards the trustor since the trustee plays a role in the trustor's belief-formation process. How the trustee presents themselves to the trustor is key for the trustor to deem the trustee trustworthy or untrustworthy. Because trust beliefs can influence the trustor's actions, they pose a potential route for manipulation that the trustee should not exploit.

When the trustee is an AI system, trust should be appropriate, meaning that the trustor's trust belief ought to be justified. The users of AI systems and individuals affected by AI's outcomes form their trust beliefs based on the information that they have of the systems. An important source of information is the image that

<sup>27</sup>For further details on the basis that lead to the conclusions in this paragraph, see section 3.2.2 and definitions 3.2, 3.4 and 3.6 in chapter 3.

the system's developers offer of their own system: what it is capable of and how it works to obtain results. With time, these trustors will obtain information from their own interactions with the system. But at first instance, the trustor's perception will be largely based on the image of (un)trustworthiness<sup>28</sup> created by the system's developers.

The trustee bears the moral responsibility of facilitating the formation of the trustor's *justified beliefs*. This means the trustee should provide clear, accurate, and accessible information, ensuring the trustor can access rational reasons to support or refute the decision to delegate a task. It is essential that the trustee does not create barriers that prevent the trustor from understanding the relevant factors. In essence, the trustee must avoid epistemically manipulating the trustor, allowing for informed and autonomous decision-making.

---

<sup>28</sup>Typically, developers promote their systems as trustworthy to ensure commercial success. However, one can picture a scenario where, despite aiming to project an image of trustworthiness, potential trustors perceive untrustworthiness instead. For example, consider Elon Musk's repeated promises about Tesla's self-driving vehicles' autonomy. While intended to portray trustworthiness, Musk's assurances have often had the opposite effect due to the promises being repeatedly delayed over the years (for details, see <https://www.bbc.com/news/business-68914929>).

---

## ***Conclusion***

---

I began this thesis driven by two central questions: can AI systems be objects of trust, and if so, which systems should be trusted and on which basis? To address these questions, I have explored the concepts of both AI and trust.

I conceptualised AI as a *socio-technical tool* (STT), meaning that it encompasses not only hardware and software but also the social dimensions in which it operates. This includes interactions between designers, policymakers, stakeholders, and users, as well as the system's technical components interacting with the data it processes. These interactions shape the system's decision-making criteria, which I interpret as the system's motives for action.

Regarding trust, I have redefined it as a *belief-based disposition* (BBD). This means that a trustor places trust in those they expect to act in a certain way, motivated by certain motives. This belief forms the basis of their willingness to delegate a task to the trusted party, under certain conditions.

Understanding AI systems as socio-technical tools whose actions can be driven by motives, AI becomes a potential object of trust. The systems that ought to be trusted are those capable of meeting human trustors' expectations while ensuring alignment with their motives. In trusting relationships, both interpersonal and with AI systems, the trustor can never be entirely sure of the trustee's future hypothetical behaviour. Trust is developed based on factors such as the trustee's past performance or evidence suggesting they are motivated by certain intentions. This inherent uncertainty places the trustor in a vulnerable position, which the trustee has a moral responsibility not to exploit. The trustee has the moral responsibility to facilitate the formation of the trustor's justified beliefs. In other words, the trustee should accurately represent their own trustworthiness to prevent the trustor from forming unrealistic expectations. This ensures that the trustor's decision to delegate a task is informed and autonomous.

The details of the above narrative have been developed throughout the six chapters of this thesis. In the first two chapters, I laid down the foundations for the rest of the thesis by clarifying core concepts related to trust, AI and, ultimately, trust in AI. Specifically, in chapter [1](#), I discussed three concepts that underpin my argument supporting the idea that trust is a suitable term for describing certain human-AI interactions: reliance, trust and trustworthiness. My aim was to provide intuitive definitions of these concepts, to unveil its nuances later in the thesis. I set the ground for this by outlining the relationship between trust and trustworthiness and the distinction between trust and reliance. In this first approach to the concept of trust, I described it as a kind of relationship. There is abundant literature on trust when applied to interpersonal or institutional relationships. However, whether the concept suits human-AI relationships is not obvious. Before continuing to deepen

the exploration of the concept of trust to determine if the latter is the case, first I needed to clarify my understanding of AI.

In chapter ??, I provided a brief introduction to AI and the use of the term in this thesis. My focus has been on ML and NNs. After briefly describing the functioning and potential of these kinds of systems, I discussed their nature as STTs. In doing so, I made it clear that when I refer to AI, I am not just addressing its technical components. On the contrary, I also take into account its social aspects, such as the involvement of humans in the design and operation of these systems, the users, and the interactions between these parties and the data they provide to the AI systems. I consider AI systems to be quasi-agents, in the sense that they can be treated as if they were agents, given their apparent autonomy. With this framework in mind, I proposed two case studies to illustrate the concept of trust in AI. I used these case studies to outline the main features to be expected in trust relationships between humans and AI, positing that trust could be extended to non-human entities. I argued that trust is characterised by six key features: the trustor's positive expectations, the (potential) wish of delegation, the risk of the trustee failing the trustor, being motives-based, being grounded in rational reasons and intersubjective. These features apply in cases of interpersonal and institutional trust as well as trust in AI. To prove this point, I provided an overview of the literature on interpersonal and institutional trust and drew parallels with my previous case studies. I concluded that the similarities were strong enough to consider trust as a suitable concept to describe some human-AI relationships. However, there are also gaps in the literature on interpersonal and institutional trust that fail to explain some specifics of trust in AI. For example, why would a trustee delegate a task to a trustor with whom they do not have a personal relationship, as is the case with artificial trustees? The inadequacy of previous accounts to describe some parts of the trusting relationship that can be fostered between a person and an AI system motivated me to develop my own account of trust.

In chapter 3, I developed the *belief-based disposition* (BBD) account of trust. Belief-based dispositions are the kind of dispositions that originate in belief. That means that belief-based dispositions are states in which an object  $x$  is disposed to  $M$  when  $C$  only under the background condition of a certain belief being held. Applied to trust, this means that the trustor trusts the trustee when the trustor is disposed to delegate some task  $\delta$  to the trustee (when the trustor needs or wishes to do so), if the trustor holds a certain belief. In the case of trust, such a belief would be a *trust-belief*. A trust-belief is the trustor's belief that the trustee is trustworthy. The trustor considers the trustee to be trustworthy if the trustee is perceived as capable of performing  $\delta$  and their motives to do  $\delta$  are approved by the trustor. The trustor has access to different kinds of reasons that shape their perception of the trustee, and ultimately lead them to believe that the trustee is more or less trustworthy. When the trustor has access to reasons that justify believing that the trustee is trustworthy, their trust is *appropriate*. As I presented them, both trust in a general sense and appropriate trust can be fostered towards people, institutions or AI systems. The goal of the BBD account was to describe trust in both human and AI trustees. After presenting the framework for understanding how the trustor relates to the trustee, I was ready to address the specifics of trust in AI.

In chapter 4, I explored the question of how to foster trust in AI systems. I focused on the black box problem as the primary obstacle to trust, even when systems appear to produce accurate results. In the literature, one of the most popular solutions to this issue is explainable AI (XAI). However, it is often overlooked that XAI will only foster trust if explanations are a suitable means for achieving that purpose. This premise has been referred to as the *Explainability-Trust Hypothesis* (Kästner et al. 2021). Building on the idea that trust is belief-based, I identified two epistemic limitations of ET:

- (a) Only explanations tailored to the epistemic background of the trustor have the potential to generate a trust-belief.
- (b) Only *good*<sup>29</sup> explanations have the potential to generate a trust-belief in the explanation recipient.

In addition, building on the BBD account presented in chapter 3, I deepened my ideas on trust-beliefs and how they are built. The BBD account allows for the conception of trust as a gradual concept, where different degrees of trust correspond to varying degrees of trust-belief. The trustor's trust will fluctuate depending on the level of justification behind their trust-belief. A trust-belief is more or less justified based on the strength of the reasons supporting it. This understanding of trust-beliefs aligns with my notion of *appropriate trust*. After discussing how trust-beliefs are formed and how they should be formed, I reaffirmed that when it comes to AI trustees, the goal should not be mere trust but appropriate trust.

Up to chapter 4, I gave a positive answer to the question of whether trust can be extended to AI trustees and began to establish criteria for determining which AI systems should be trusted. The first part of this effort focused on epistemic criteria. In the remainder of the thesis, I approached the question from an ethical perspective.

Chapter 5 set the groundwork for the final claim of this thesis by addressing potential objections, reinforcing the argument that trust possesses a moral dimension which makes it a fitting concept for describing human-AI relationships. I discussed papers by Nickel, Franssen, and Kroes (2010), Ryan (2020), and Durán and Jongsma (2021). These papers share their opposition to the idea that non-human entities such as AI systems can be trusted. Instead, these authors argue that AI systems could be at best relied on, proposing different forms of reliance that fit the context of human-AI interaction better, according to them. I analysed their arguments in an attempt to address potential criticisms against my BBD account of trust, and the following claims that were to follow in the next chapter. I used the BBD account of trust and the characterization of trust, reliance and motives presented in chapters 1 and chapter 2 as a basis to argue that reliance does not suffice to describe the kind of relationships that would be desirable to have in some human-AI interactions.

Finally, in chapter 6, I examined the moral dimension of trust, particularly in relation to AI. The central takeaway is that the trustee, whether human or artificial, bears the moral responsibility of accurately representing their trustworthiness to the

---

<sup>29</sup>What is a good explanation constitutes a rich debate in epistemology. The details of this debate go beyond the scope of this thesis. However, in chapter 6, I will attempt to at least clarify what I am referring to by *good explanations* in the context of XAI.

trustor. In the case of AI, this responsibility should be assumed by the elements of the system capable of accountability –namely, human agents such as designers or stakeholders. I argued that trust, unlike mere reliance, is a more appropriate framework for high-stakes human-AI interactions due to its moral dimension, where trustees have a responsibility to ensure they do not mislead trustors about their capabilities or intentions. This responsibility is particularly important in situations where the trustor is vulnerable and lacks the background to fully understand the AI’s decision-making processes. By integrating these moral considerations into my belief-based disposition (BBD) account of trust, I have provided a framework for understanding trust in AI, alongside its ethical implications, offering a comprehensive approach to trust in human-AI interactions.

In conclusion, this thesis has demonstrated that trust, rather than mere reliance, is a more fitting framework for understanding and guiding human-AI relationships, especially in high-stakes contexts. By conceptualising AI as STTs and integrating the moral dimension of trust into the BBD account, I have shown that trust in AI is not only possible but also normatively desirable. The moral responsibility of trustees –whether human or artificial– to accurately represent their trustworthiness to trustors is crucial to fostering appropriate trust. This framework offers a solid foundation for discussing trust in AI, rather than relying on superficial analogies to interpersonal relationships. The way we conceptualise and speak about human-AI interactions directly shapes those relationships and, in turn, AI’s integration into society. Establishing a coherent and well-grounded discourse on trust in AI not only contributes to a deeper understanding of these relationships but also highlights the broader implications of how AI systems are perceived and trusted. As AI systems continue to evolve and play a more significant role in our lives, cultivating appropriate trust in them will be essential for both technological development and societal well-being.

**A**

---

*Trust: Related Concepts*

---

---

## A.1 Distrust

It is uncontroversial to say that trust and distrust are contraries but not contradictories (Jones [1998]). When characterizing trust, it is pertinent to define distrust too in order to fully understand it and to avoid confusion with other concepts such as mistrust or failed trust. Quoting Hawley ([2014]), “distrust is not mere absence of trust” (p. 3). A trustor distrusts a trustee when they think the trustee would fail to execute the task which they could hypothetically be trusted with. For example, I distrust my neighbour to take care of my plants if I rather would not ask them to do so, because I think that if I did my plants would see the end of their days soon. This does not mean that in the case that my intention was to kill my plants, I would ask my neighbour to take care of them because I know she would fail and I would achieve what I wanted. Distrust does not mean either that my neighbour has already failed to take care of my plants and now I regret having trusted her in the first place. This example shows that trust is not a coin of two sides being the other one distrust, but a much richer concept.

As I have already noted, trusting comes with a certain degree of vulnerability. The trustor makes themselves vulnerable to the risk of the failure of the trustee. Even though we expect them to perform the trusted task successfully, it always exists the possibility of them not doing so. This applies both to human and artificial trustees. It is important to be aware that trusting AI does not mean to believe that the system is bulletproof. The risk of failure is always there, but AI could be trusted when there are reasons to believe that such risk is unlikely enough to actually be the case. In contrast, when the possibility of failure becomes a likely scenario in the eyes of the trustor, distrust arises.

---

## A.2 Mistrust

Distrust is the lack of conviction of the trustor that the trustee is trustworthy. This absence of trust does not need to be due to any particular motive. I do not trust my neighbour with my plants for no particular reason; I just do not know her well enough, so I distrust her. But sometimes, one can have active reasons not to trust someone. I do know my ex-boyfriend quite well, and precisely because I know him I would never trust him with my plants. He has an extensive record of deceased monstera and that is exactly the reason why I won’t ask him to take care of mine. In this case, I do not merely distrust my ex-boyfriend but I *mistrust* him.

Bringing back the distinction between trust and reliance made in section [1.6], one could think that mistrusting someone means simply finding them unreliable. Perceived unreliance definitively leads to mistrust. However, they are not interchangeable concepts since mistrust can arise due to other reasons. Maybe my ex-boyfriend is incapable of taking care of plants (therefore he is unreliable), or maybe he does

have the skill but what he does not have is the will of doing me any kind of favour. In both cases, I would mistrust him.

Mistrust refers to the absence of trust towards the trustee because of specific reasons. It is not only that the trustor is uncertain whether the trustee will succeed or not in the trusted task so they would rather not trust them. The trustor feels certain about the trustee's failure because they have reasons not to delegate on them. In this sense, mistrust is stronger than distrust.

---

### A.3 Failed trust

So far I have covered the scenarios in which the trustee had either no reasons to trust (therefore they distrust) or active reasons to mistrust. A different kind of scenario regards the cases in which the trustor is overly sceptical, so they distrust the trustee, even though the latter is actually trustworthy. I do not trust my neighbour to water my plants because I know nothing about them, so I have no reason to believe she will make a good job. But maybe my neighbour turns out to be a secret gardener and there could not be a better candidate for me to trust. So I should have trusted my neighbour, but I did not. This is what is called failed trust (Saultz 2020).

Failing to trust trustworthy trustees is undesirable. Maybe not as much as mistakenly trusting untrustworthy ones, but still, it poses a situation to be avoided. Trustors have the responsibility to place our trust correctly and successfully distinguishing (to the best of our abilities) untrustworthy trustees from trustworthy ones (Hawley 2014). That is how (dis)trust plays a valuable role in society, both by allowing us to delegate and achieve higher goals and by protecting us from the betrayal of untrustworthy trustees.

---

### A.4 Misplaced trust

Finally, I briefly point at cases in which the trustor thought the trustee to be trustworthy when they were not. Following the go-to example of this section, before mistrusting my ex-boyfriend, I misplaced my trust in him (and that is why I mistrust him now). My trust in him was misplaced because I did trust him to take care of my plants once, and he failed miserably. He portrayed a fake image of trustworthiness, which led me to (unwarrantedly) trust him. Cases of misplaced trust take place often because as stated in section 1.5, trust and trustworthiness are independent from each other (even though they should not be).



# B

---

## *Dispositions and Beliefs*

---

## B.1 Dipositions

### B.1.1 Brief history of dispositions

Dispositions are special kind of properties that have raised philosophical interest since centuries ago. Part of what makes them so special is that, despite undeniable, they are not observable. A proposition such as “glass is fragile” has a clear intuitive meaning: glass is easy to break, easier than most materials. I know this because it is part of my mundane experience, and I get a reminder every time someone drops their drink at a local pub, which, after midnight, tends to happen often enough. However, while the drink is still on the table, there is no indicator that immediately exhibits the glass’ fragility. I know that the glass will break, *if it falls*. But until it does, it just lays on the table unyielding, together with objects made of more sturdy materials, such as the ashtray or my phone. Dispositions refer to properties of the object that would make it behave in a certain way under certain circumstances. Here lies the difference with other properties such as “redness”. If the ashtray is red, there is no need for an event that triggers such a property. That is how the ashtray is, and I continuously perceive it this way. This kind of properties, by opposition to disposition, are called categorical properties. But before digging deeper into the distinction, let me dive into a brief historical overview of the concept of disposition.

In his *Metaphysics*, Aristotle (1925, book IX) introduces the terms *energeia* and *dynamis*, referring to what it currently is and what has the potentiality of eventually being, respectively. By this distinction, he brings to light the contrast between current and potential states, referring the latter to the inherent capacity for an object to develop into a different state. Aristotle’s notion of *dynamis* could be understood as a wider way to see dispositions, since in Aristotle, the focus is on the world’s changing nature. For him, whilst an object could evolve into something else, there was some intrinsic property within the object that allowed this change to happen. These properties represented the inherent potentials or capacities of objects to change or act in certain ways under the right conditions. For example, seeds have the potentiality to become grown plants under suitable conditions.

In the Enlightenment, Locke (1812) (II.viii.8) distinguishes between *primary*, *secondary* and *tertiary qualities*. The *primary qualities* are the inseparable features of a body, like its extension. *Secondary qualities* do not exist in objects themselves but have the power to produce some sort of idea in us as experiencers; an example of secondary quality is colour, whilst primary qualities of an object –such as its matter– interact with light in such a way that makes an external observer perceive the object’s colour. Lastly, *tertiary qualities* refer to the power to alter other bodies’ primary qualities; for example, water has the power to dissolve instant coffee, being this power a tertiary quality of the water. It is interesting how Locke’s tertiary qualities attribute active power to some objects to change others. This can be seen as a complementary to the contemporaneous notion of disposition, according to which the passive object is disposed to be changed.

Locke’s early notion of disposition is contradicted by Leibniz (2006) (II, 8, § 9), who blurs out the distinction between primary and tertiary qualities. According

to him, “intelligible powers” would be a tertiary quality (as opposed to Locke, who considered them as a primary quality). In Leibniz’s writings, “intelligible powers” refer to the inherent capacities or abilities possessed by substances, especially monads, to produce effects or undergo changes in a way that is intelligible or rationally explicable. In addition, Leibniz’s concept of *force* (not to be confused with the contemporary notion of force in physics) is considered by some as an alternative way to understand dispositions (see Liske (2009)).

In sum, throughout history, different authors have considered different kinds of properties to be dispositions. While properties such as “redness” had been considered as “the disposition to be perceived as red” by some authors (Johnston 1992, p. 225)<sup>1</sup>, today it is widely accepted that properties designating colour or extension are categorical, while those which typically need a trigger to manifest a change of state are dispositional (Mumford 1998, pp. 20–22; Choi and Fara 2012).

It is relatively uncontroversial to say that a disposition is a property characterized by having a *stimulus condition* and a *manifestation* (Choi and Fara 2012). In a disposition, the stimulus is the trigger for a specific behaviour. Such behaviour is the manifestation. Going back to the initial fragility example: the stimulus condition for a fragile object  $x$  would be to be struck (hard enough)<sup>2</sup> and the manifestation, to break. Thus,  $x$  is fragile if, when struck,  $x$  breaks. Another classic example is solubility: in this case, the stimulus condition is putting an object  $x$  in water, and the manifestation is  $x$  being dissolved.

Thus, dispositions refer not only to the current state of the world but to states of the world in which some condition obtains, such as  $x$  being put in water. That is why dispositions are commonly analysed in terms of counterfactual conditionals, which are statements of the sort “if  $X$  had been...  $Y$  would have been...” (Lewis 1973). This takes me back to the definition of disposition offered in chapter 3:<sup>3</sup>

**Definition B.1.** A disposition is a property  $D$ , of an object  $x$ , that is characterized by an associated stimulus condition  $C$  and manifestation  $M$ . The object  $x$  possesses  $D$  iff  $x$  would  $M$  if it were the case that  $C$ .

The definition above does not come without problems. One that is particularly relevant for this thesis, is whether the stimulus condition  $C$  must be a single condition associated with the manifestation  $M$ , or there can be multiple stimulus conditions  $C_1, C_2, \dots$  that trigger  $M$ . The latter concept is known as *multi-track disposition*.

## B.1.2 Single versus multi-track dispositions

The concept of multi-track dispositions refers to the idea that the manifestation of a dispositional property could be triggered by different stimulus conditions (Ryle 1949, 43–45). Going back to the initial fragility example, being struck is not the

<sup>1</sup>Johnston uses as an example canary yellow instead of red; I have opted for the latter colour simply because the word “yellowness” does not resonate with me in quite the same manner.

<sup>2</sup>The need for this parenthesis already hints at the difficulty to pin down stimulus conditions. While the concept of fragility may appear straightforward initially, attempting to explicitly identify the characteristics that render an object fragile quickly exposes the complexity of this undertaking.

<sup>3</sup>Definition B.1 reflects the so-called Simple Conditional Analysis (SCA) of dispositions. Further details can be found in the main text in section 3.3.2

only condition under which the fragility of my glass could be observed. The glass could be dropped, stressed, filled with boiling water, etc., being the result always the same: my glass being broken and me having no drink. One shall not confuse multi-track dispositions with ill-defined single-track dispositions; that is, classical dispositions with a unique stimulus condition that has been misleadingly split into several context-dependant cases. For example, it would be wrong to label the disposition to be poisonous as multi-track due to the fact that a substance  $x$  is poisonous when ingested by humans, by dogs, etc. (Choi and Fara 2012).

Some authors (Cartwright 1999, pp. 59–64; Vetter 2013) consider that, rather than special cases, multi-track dispositions are the norm: for each disposition  $D$ , there is a variety of stimulus conditions  $C_1, C_2, \dots$  that trigger  $M$ . An illustration of this perspective is seen in Vetter’s argument, where she goes as far as asserting that for each disposition, there is no privileged subset of stimulus conditions per disposition. Consequently, she does not regard stimulus conditions as characteristic of the concept of disposition, but rather focus on the manifestation (Vetter 2014, p. 132).

Under the assumption that (at least some) dispositions are multi-track, it has been argued that beliefs are dispositional. That takes me to the second and last part of this appendix.

## B.2 Beliefs

### B.2.1 Beliefs as multitrack dispositions

Quoting Armstrong (1973, p. 8), since Ryle (1949), the comparison between beliefs and dispositions has been common in philosophy. Picking up the argument where I left off in the previous section, it is not uncontroversial to say that dispositions can have multiple stimulus manifestations and manifestations. My glass can be damaged in different ways, resulting in the glass being broken or cracked. Parallely, some authors (Armstrong 1973, pp. 8–9; Ryle 1949, pp. 109–131) argue that beliefs can be manifested in several ways, such as acts of speech. For example, my belief that my glass is transparent can be manifested by my out loud proclamation that my glass is transparent. However, even if I keep this idea to myself, I still believe that the glass is transparent. This is what has been referred as *dispositional belief*.

### B.2.2 Dispositional beliefs

The term “dispositional belief” can refer to beliefs that an agent holds dormantly in the background of their mind, in contrast to occurrent beliefs that are currently being entertained (Blaauw 2013, pp. 90–91). For example, I believe that the Earth is round, but I was not recalling such a belief to my mental forefront until I started thinking of an example of a trivial belief that I have no need to focus on that often. In what pertains to this thesis, I am more interested in the use of “dispositional belief” which followed Ryle’s work.

In *The Concept of the Mind*, Ryle (1949) proposed the concept of “dispositions”

as a way to understand mental states. According to him, mental states are not abruptly separated from observable external reality, but manifest themselves in such a reality. In this sense, agents' behaviour is the manifestation of their mental states, including their beliefs. This fits the characterization above of dispositions as multitrack: multiple stimulus conditions can make a single belief manifest in multiple manners. For example, consider once again that I believe that my glass is transparent. I can manifest such a belief triggered by different conditions: that you ask me about it, an awkward silence that I try to overcome, my need to let my inner monologue out... there are also different manifestations that can make my belief observable. For example, instead of manifesting my belief by verbalizing it, I could manifest it by pouring wine into the glass not as generously as I wished, since I believe that the glass is transparent, therefore you can see you much wine I intend to drink and I act accordingly to not feel judged.

The aim of this appendix is to distinguish the idea above from the concept of belief-based disposition (BBD) that I proposed in chapter 3. Unlike dispositional beliefs, BBDs refer to states that *originate* in belief. That is, an object  $x$  possess a certain BBD iff  $x$  is disposed to  $M$  when  $C$  only under the background condition of a certain belief being held. In the case of trust, trust belief is the kind of belief that is required for trust to occur. Thus, the concept of BBD is a little more complicated than dispositional beliefs since belief is only one of two conditions that need to be satisfied. Let me briefly revisit definition 3.4:

**Definition B.2.** *Trust* is the belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold

- (i)  $A$  has a continuing belief that  $B$  is trustworthy.
- (ii) If  $A$  wished or needed  $B$  to do  $\delta$ , then  $A$  would be willing to delegate  $\delta$  to  $B$ .

The dispositional aspect of the definition above is described in condition (ii). Condition (i) is an additional condition that needs to be met for trust to occur, rather than (ii) being the manifestation of the trust belief referred to in (i). If a trustor  $A$  holds a trust belief but (ii) is not satisfied,  $A$  does not really trust  $B$ . Therefore, trust is not just a dispositional belief which finds its manifestation in delegation. In consequence, I find myself in the need to introduce the concept of belief-disposition, to which chapter 3 is devoted.



---

## *Glossary*

---

**Appropriate trust** is the *justified* belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold: (i')  $A$  has a continuing justified belief that  $B$  is trustworthy, and (ii'') if  $A$  wished or needed  $B$  to do  $\delta$ , then  $A$  would be willing to delegate  $\delta$  to  $B$ . [72](#), [135](#)

**Belief-based disposition** refers to the kind of dispositions that originate in belief. That means that belief-based dispositions are states in which an object  $x$  is disposed to  $M$  when  $C$  only under the background condition of a certain belief being held. [71](#)

**Disposition** refers to a kind of property  $D$ , of an object  $x$ , that is characterized by an associated stimulus condition  $C$  and manifestation  $M$ . The object  $x$  possesses  $D$  iff  $x$  would  $M$  if it were the case that  $C$ . [69](#)

**Explainability-Trust Hypothesis** (ET\*) is the claim that explanations are a suitable means for facilitating trust in a potential trustor  $A$ , if and only if the explanation has the potential to generate in  $A$  a trust belief. To generate a trust belief, the explanation must be of good quality and tailored to  $A$ 's particular doxastic disposition. Such a disposition includes  $A$ 's background knowledge and the effect of the explanation on  $A$ . [88](#), [97](#)

**Motive** refers to the criterion that favours a certain course of action. In the case of humans, such criterion is the combination of desire and belief that drives an agent's actions. That is, what a person believes and desires conform their standard to decide what to do. For example, I believe that cats are cute and I desire to spend time surrounded by them, which acts as a motive for me to take care of my neighbour's cat when I have the chance to do so. When it comes to artificial quasi-agents such as AI systems, the criterion that favours the course of action is not conformed by beliefs and desires since they have neither of those. Instead, AIs actions are favoured by the criterion they use to minimize a loss function. A loss function is a mathematical representation of the difference between the system's predictions and the actual outcomes. During the training process, a NN iteratively adjusts its internal parameters (often referred to as weights) to minimize this loss function. The process that the NN uses to adjust its weights and 'learn' over time is what guides its actions. In this sense, I consider these processes the AI's motives. [50](#), [127](#)

**Reliance** is the dependence that one party has towards another when the first needs to delegate something to the second. [13](#)

**Socio-technical tool (STT)** refers to systems which consist of human and non-human agents who interact for various purposes. I use the term “socio-technical tool” instead of “socio-technical system” because, to my understanding, a socio-technical system comprehends a higher degree of complexity, given that includes wider processes than the ones I aim to point at (see the characterization of socio-technical systems offered by Van House (2003)). For example, a book publishing system would be a socio-technical system that includes authors, publishing companies, printers, data-bases, etc. By “socio-technical tool”, I refer to specific models and devices, such as OpenAI’s ChatGPT (Generative Pre-trained Transformer), for example. [27](#)

**Trust** is the belief-based disposition of a trustor  $A$  to delegate some task or role  $\delta$  (belonging to a domain  $\Delta$ ) to a trustee  $B$ . That means that  $A$  trusts  $B$  iff both of the following hold: (i)  $A$  has a continuous belief that  $B$  is trustworthy, and (ii) if  $A$  wanted or needed  $B$  to do a certain task for them, then  $A$  would be willing to delegate that task to  $B$ . While both trust and reliance potentially lead to delegation, trust is a stronger concept. When trusting, the trustor believes that the trustee would most likely succeed in doing what is delegated to them. Conversely, it is possible to rely on someone without believing that they will successfully do whatever we delegate to them (but still relying on them due to a lack of alternatives). [15](#), [67](#), [70](#), [100](#), [149](#)

**Trustworthiness** is the relational property of a trustee  $B$  being worthy of a trustor  $A$ ’s trust, in some particular domain  $\Delta$ .  $B$  is worthy of  $A$ ’s trust (regarding  $\Delta$ ) iff it is reasonable for the trustor to delegate some task or role  $\delta$  (belonging to  $\Delta$ ) to them. It is reasonable for  $A$  to delegate  $\delta$  to  $B$  iff  $A$  has access to rational reasons that lead to the conclusion that  $B$  is likely to perform  $\delta$  successfully, moved by motives deemed appropriate by  $A$ . [17](#), [67](#), [100](#)

---

## ***Bibliography***

---

- Adadi, Amina and Mohammed Berrada (2018). “Peeking Inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6, pp. 52138–52160. DOI: [10.1109/ACCESS.2018.2870052](https://doi.org/10.1109/ACCESS.2018.2870052).
- Adler, Jonathan E. (1994). “Testimony, Trust, Knowing”. In: *The Journal of Philosophy* 91.5, pp. 264–275. DOI: [10.2307/2940754](https://doi.org/10.2307/2940754).
- Ågerfalk, Pär J (2020). “Artificial Intelligence as Digital Agency”. In: *European Journal of Information Systems* 29.1, pp. 1–8. DOI: [10.1080/0960085X.2020.1721947](https://doi.org/10.1080/0960085X.2020.1721947).
- Alvarado, Ramón (2023). “What Kind of Trust does AI Deserve, if Any?” In: *AI and Ethics* 3.4, pp. 1169–1183. DOI: [10.1007/s43681-022-00224-x](https://doi.org/10.1007/s43681-022-00224-x).
- Ananny, Mike (2016). “Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness”. In: *Science, Technology, & Human Values* 41.1, pp. 93–117. DOI: [10.1177/0162243915606523](https://doi.org/10.1177/0162243915606523).
- Anderson, Elizabeth (2012). “Epistemic Justice as a Virtue of Social Institutions”. In: *Social Epistemology* 26.2, pp. 163–173. DOI: [10.1080/02691728.2011.652211](https://doi.org/10.1080/02691728.2011.652211).
- Anjomshoae, Sule et al. (2019). “Explainable Agents and Robots: Results from a Systematic Literature Review”. In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pp. 1078–1088. ISSN: 978-1-4503-6309-9.
- Anscombe, Gertrude Elizabeth Margaret (1957). *Intention*. Basil Blackwell.
- Aristotle (1925). *Metaphysics: Book IX*. (W.D. Ross, Trans.): The Internet Classics Archive, (Original work published 350 B.C.E). URL: <http://classics.mit.edu/Aristotle/metaphysics.9.ix.html>.
- Armstrong, David Malet (1973). *Belief, Truth and Knowledge*. Cambridge University Press.
- Bachmann, Reinhard (2020). “Trust and Institutions”. In: *Encyclopedia of Business and Professional Ethics*. Ed. by Deborah C Poff and Alex C Michalos. Cham: Springer International Publishing, pp. 1–6. ISBN: 978-3-319-23514-1.
- Baier, Annette (1986). “Trust and Antitrust”. In: *Ethics* 96.2, pp. 231–260. DOI: [10.1086/292745](https://doi.org/10.1086/292745).
- Baker, Judith (1987). “Trust and Rationality”. In: *Pacific Philosophical Quarterly* 68.1, pp. 1–13. DOI: [10.1111/j.1468-0114.1987.tb00280.x](https://doi.org/10.1111/j.1468-0114.1987.tb00280.x).
- Barad, Karen (1998). “Getting Real: Technoscientific Practices and the Materialization of Reality”. In: *Differences* 10.2, pp. 87–128. DOI: [10.1215/10407391-10-2-87](https://doi.org/10.1215/10407391-10-2-87).
- Bellia, Anthony J. (2002). “Promises, Trust and Contract Law”. In: *The American Journal of Jurisprudence* 47, pp. 25–40. DOI: [10.1093/ajj/47.1.25](https://doi.org/10.1093/ajj/47.1.25).
- Bellman, Richard Ernest (1978). *Introduction to Artificial Intelligence. Can Computers Think?* San Francisco: Boyd & Fraser.

- Benk, Michaela et al. (2022). “The Value of Measuring Trust in AI: A Socio-Technical System Perspective”. In: *CHI 2022-Workshop on Trust and Reliance in AI-Human Teams (TRAIT)*, pp. 1–12. DOI: [10.48550/arXiv.2204.13480](https://doi.org/10.48550/arXiv.2204.13480).
- Bird, Alexander (1998). “Dispositions and Antidotes”. In: *The Philosophical Quarterly* 48.191, pp. 227–234. ISSN: 0031-8094.
- Blaauw, Martijn (2013). “Contrastive belief”. In: *Contrastivism in Philosophy* 39, pp. 88–100. ISSN: 0415878608.
- Blanco, Sara (2022). “Trust and Explainable AI: Promises and Limitations”. In: *EthiComp*. Turku, pp. 245–256. ISBN: 978-951-29-8989-8. URL: <https://sites.utu.fi/ethiComp2022/wp-content/uploads/sites/1104/2022/09/EthiComp-2022-Proceedings-Corrected.pdf>.
- Bordt, Sebastian et al. (2022). “Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency* 1.1, pp. 891–905. DOI: [10.1145/3531146.3533153](https://doi.org/10.1145/3531146.3533153).
- Buchholz, Oliver (2023). “A Means-End Account of Explainable Artificial Intelligence”. In: *Synthese* 202.33, pp. 1–23. DOI: [10.1007/s11229-023-04260-w](https://doi.org/10.1007/s11229-023-04260-w).
- Buijsman, Stefan (2022). “Defining Explanation and Explanatory Depth in XAI”. In: *Minds and Machines* 32.3, pp. 563–584. DOI: [10.1007/s11023-022-09607-9](https://doi.org/10.1007/s11023-022-09607-9).
- Burrell, Jenna (2016). “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms”. In: *Big Data and Society* 3.1, pp. 1–12.
- Callon, Michel (2004). “Actor Network Theory”. In: *International Encyclopedia of the Social & Behavioral Sciences*. Ed. by Neil J Smelser and Paul B Bates. Amsterdam: Elsevier, pp. 62–66.
- Cartwright, Nancy (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press. ISBN: 0521644119.
- Castelfranchi, Christiano and Rino Falcone (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. Sussex: John Wiley & Sons.
- Choi, Shungho and Michael Fara (2012). “Dispositions”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N Zalta. Spring 2021 Edition. URL: <https://plato.stanford.edu/archives/spr2021/entries/dispositions/>.
- Coeckelbergh, Mark (2009). “Virtual Moral Agency, Virtual Moral Responsibility: On the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents”. In: *AI & Society* 24.2, pp. 181–189. DOI: [10.1007/s00146-009-0208-3](https://doi.org/10.1007/s00146-009-0208-3).
- (2012). “Can We Trust Robots?” In: *Ethics and Information Technology* 14.1, pp. 53–60. DOI: [10.1007/s10676-011-9279-1](https://doi.org/10.1007/s10676-011-9279-1).
- (2019). “Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability”. In: *Science and Engineering Ethics* 26.4, pp. 2051–2068. DOI: [10.1007/s11948-019-00146-8](https://doi.org/10.1007/s11948-019-00146-8).
- (2020). *AI Ethics*. Cambridge: The MIT Press.
- Cogley, Zac (2012). “Trust and the Trickster Problem”. In: *Analytic Philosophy* 53.1, pp. 30–47. DOI: [10.1111/j.2153-960X.2012.00546.x](https://doi.org/10.1111/j.2153-960X.2012.00546.x).
- Coleman, James S (1990). *Foundations of Social Theory*. Cambridge: Harvard University Press.
- Cook, Karen S, Russell Hardin, and Margaret Levi (2005). *Cooperation Without Trust?* New York: Russell Sage Foundation.

- Creel, Kathleen A. (2020). “Transparency in Complex Computational Systems”. In: *Philosophy of Science* 87.4, pp. 568–589. DOI: [10.1086/709729](https://doi.org/10.1086/709729).
- Davidson, Donald (1963). “Actions, Reasons, and Causes”. In: *Journal of Philosophy* 60.23, pp. 685–700. DOI: [10.2307/2023177](https://doi.org/10.2307/2023177).
- Dennett, Daniel C. (1988). *The Intentional Stance*. Cambridge: The MIT Press. DOI: [10.1017/S0140525X00058611](https://doi.org/10.1017/S0140525X00058611).
- Deutsch, David (2009). *A New Way to Explain Explanation*. URL: [https://www.ted.com/talks/david\\_deutsch\\_a\\_new\\_way\\_to\\_explain\\_explanation](https://www.ted.com/talks/david_deutsch_a_new_way_to_explain_explanation).
- Dignum, Virginia et al. (2018). “Ethics by Design: Necessity or Curse?” In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 60–66. DOI: [10.1145/3278721.3278745](https://doi.org/10.1145/3278721.3278745).
- Durán, Juan Manuel and Nico Formanek (2018). “Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism”. In: *Minds and Machines* 28, pp. 645–666. DOI: [10.1007/s11023-018-9481-6](https://doi.org/10.1007/s11023-018-9481-6).
- Durán, Juan Manuel and Karin Rolanda Jongsma (2021). “Who is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI”. In: *Journal of Medical Ethics* 47.5, pp. 329–335. DOI: [10.1136/medethics-2020-106820](https://doi.org/10.1136/medethics-2020-106820).
- Elgin, Catherine Z (2005). “Non-Foundationalist Epistemology: Holism, Coherence, and Tenability”. In: *Contemporary Debates in Epistemology*. Ed. by Mattias Steup and Ernest Sosa. Malden: Blackwell Publishing, pp. 156–167.
- Faulkner, Paul (2015). “The Attitude of Trust is Basic”. In: *Analysis* 75.3, pp. 424–429. DOI: [10.1093/analys/anv037](https://doi.org/10.1093/analys/anv037).
- Ferrario, Andrea, Michele Loi, and Eleonora Viganò (2020). “In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions”. In: *Philosophy and Technology* 33.3, pp. 523–539. DOI: [10.1007/s13347-019-00378-3](https://doi.org/10.1007/s13347-019-00378-3).
- Floridi, Luciano et al. (2018). “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”. In: *Minds and Machines* 28.4, pp. 689–707. DOI: [10.1007/978-3-030-81907-1/3](https://doi.org/10.1007/978-3-030-81907-1/3).
- Foley, Richard (2001). *Intellectual Trust in Oneself and Others*. New York: Cambridge University Press.
- Fox-Decent, Evan (2005). “The Fiduciary Nature of State Legal Authority”. In: *Queen’s LJ* 31, pp. 259–310. URL: <https://ssrn.com/abstract=1090292>.
- Frankfurt, Harry (1971). “Freedom of the Will and the Concept of a Person”. In: *Journal of Philosophy* 1.68, pp. 5–20. DOI: [10.2307/2024717](https://doi.org/10.2307/2024717).
- Fricker, Miranda (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.
- (2010). “Can There Be Institutional Virtues?” In: *Oxford Studies in Epistemology*. Vol. 3. Oxford: Oxford University Press, pp. 235–252.
- Fried, Charles (2015). *Contract as a Promise: Theory of Contractual Obligation*. Oxford: Oxford Scholarship Online.
- Gambetta, Diego (1988). “Can We Trust Trust?” In: *Trust. Making and breaking cooperative relations*. Ed. by Diego Gambetta. Oxford: Basil Blackwell. Chap. 13, pp. 213–237.

- Gärdenfors, Peter (1988). *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press.
- Gilpin, L H et al. (2018). “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. DOI: [10.1109/DSAA.2018.00018](https://doi.org/10.1109/DSAA.2018.00018).
- Goebel, Randy et al. (2018). “Explainable AI: The New 42?” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11015 LNCS, pp. 295–303. DOI: [10.1007/978-3-319-99740-7\\_{\\\_}21](https://doi.org/10.1007/978-3-319-99740-7_{\_}21).
- Goldberg, Sanford C. (2020). “Trust and Reliance”. In: *The Routledge Handbook of Trust and Philosophy*. Ed. by Judith Simon. New York: Routledge, pp. 97–108.
- Goodman, Nelson (1954). *Fact, Fiction and Forecast*. Cambridge, Mass: Harvard University Press.
- Guidotti, Riccardo et al. (2018). “A Survey of Methods for Explaining Black Box Models”. In: *ACM Comput. Surv* 51.93, pp. 1–42. DOI: [10.1145/3236009](https://doi.org/10.1145/3236009).
- Gunning, David et al. (2019). “XAI: Explainable Artificial Intelligence”. In: *Science Robotics* 4.37. DOI: [10.1126/scirobotics.aay7120](https://doi.org/10.1126/scirobotics.aay7120).
- Halpern, Joseph Yehuda (2003). *Reasoning About Uncertainty*. Cambridge: The MIT Press.
- Hardin, Russell (1993). “The Street-Level Epistemology of Trust”. In: *Politics & Society* 21.4, pp. 505–529. DOI: [10.1515/auk-1992-0204](https://doi.org/10.1515/auk-1992-0204).
- (2002). *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- Hawley, Katherine (2014). “Trust, Distrust and Commitment”. In: *Nous* 48.1, pp. 1–20. DOI: [10.1111/nous.12000](https://doi.org/10.1111/nous.12000).
- (2017). “Trustworthy Groups and Organizations”. In: *The Philosophy of Trust*. Oxford: Oxford University Press. ISBN: 9780198732549.
- Heaton, Dan et al. (2024). “‘ChatGPT Says No’: Agency, Trust, and Blame in Twitter Discourses after the Launch of ChatGPT”. In: *AI and Ethics*, pp. 1–23. DOI: [10.1007/s43681-023-00414-1](https://doi.org/10.1007/s43681-023-00414-1).
- Hieronymi, Pamela (2008). “The Reasons of Trust”. In: *Australasian Journal of Philosophy* 86.2, pp. 213–236. DOI: [10.1080/00048400801886496](https://doi.org/10.1080/00048400801886496).
- HLEG (2019). *Ethics Guidelines for Trustworthy AI*. URL: <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.
- Hoffman, Robert R, Gary Klein, and Shane T Mueller (2018). “Explaining Explanation For ‘Explainable AI’”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62.1, pp. 197–201. DOI: [10.1177/1541931218621047](https://doi.org/10.1177/1541931218621047).
- Hoffman, Robert R et al. (2018). *Metrics for Explainable AI: Challenges and Prospects*. Tech. rep. Institute for Human and Machine Cognition. URL: <https://arxiv.org/abs/1812.04608>.
- Holton, Richard (1994). “Deciding to Trust, Coming to Believe”. In: *Australasian Journal of Philosophy* 72.1, pp. 63–76. DOI: [10.1080/00048409412345881](https://doi.org/10.1080/00048409412345881).
- IBM Cloud Education (2020). *Neural Networks*. URL: <https://www.ibm.com/uk-en/cloud/learn/neural-networks>.

- Jacovi, Alon and Yoav Goldberg (2020). “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205. DOI: [10.48550/arXiv.2004.03685](https://doi.org/10.48550/arXiv.2004.03685).
- Jacovi, Alon et al. (2021). “Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI”. In: *FACCT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635. DOI: [10.1145/3442188.3445923](https://doi.org/10.1145/3442188.3445923).
- Johnston, Mark (1992). “How to Speak of the Colors”. In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 68.3, pp. 221–263. ISSN: 0031-8116.
- Jones, Andrew J I, Alexander Artikis, and Jeremy Pitt (2013). “The Design of Intelligent Socio-Technical Systems”. In: *Artificial Intelligence Review* 39.1, pp. 5–20. DOI: [10.1007/s10462-012-9387-2](https://doi.org/10.1007/s10462-012-9387-2).
- Jones, Karen (1996). “Trust as an Affective Attitude”. In: *Ethics* 107.1, pp. 4–25. DOI: [10.1086/233694](https://doi.org/10.1086/233694).
- (1998). “Trust”. In: *Routledge Encyclopedia of Philosophy*. Taylor and Francis. DOI: [10.4324/9780415249126-L107-1](https://doi.org/10.4324/9780415249126-L107-1).
- (1999). “Second-Hand Moral Knowledge”. In: *The Journal of Philosophy* 96.2, pp. 55–78. ISSN: 0022362X. DOI: [10.2307/2564672](https://doi.org/10.2307/2564672).
- Kappel, K (2014). “Believing on Trust”. In: *Synthese* 191, pp. 2009–2028. DOI: [10.1007/s11229-013-0376-z](https://doi.org/10.1007/s11229-013-0376-z).
- Kästner, Lena et al. (2021). “On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness”. In: *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 169–175. DOI: [10.1109/REW53955.2021.00031](https://doi.org/10.1109/REW53955.2021.00031).
- Kattukaran, Anish (2024). *Your Smart Home Is Getting Smarter, with Help from Gemini*. URL: <https://blog.google/products/google-nest/gemini-google-home/>.
- Keren, Arnon (2014). “Trust and Belief: A Preemptive Reasons Account”. In: *Synthese* 191, pp. 2593–2615. DOI: [10.1007/s11229-014-0416-3](https://doi.org/10.1007/s11229-014-0416-3).
- (2020). “Trust and Belief”. In: *The Routledge Handbook of Trust and Philosophy*. Ed. by Judith Simon. New York: Routledge, pp. 109–120.
- Kiener, Maximilian (2022). “Can We Bridge AI’s Responsibility Gap at Will?” In: *Ethical Theory and Moral Practice* 25.4, pp. 575–593. DOI: [10.1007/s10677-022-10313-9](https://doi.org/10.1007/s10677-022-10313-9).
- Kirby, Nicolas, Andrew Kirton, and Aisling Crean (2018). “Do Corporations Have a Duty to Be Trustworthy?” In: *Journal of the British Academy* 6.s1, pp. 75–129. DOI: [10.5871/jba/006s1.075](https://doi.org/10.5871/jba/006s1.075).
- Kirton, Andrew (2020). “Matters of Trust as Matters of Attachment Security”. In: *International Journal of Philosophical Studies* 28.5, pp. 583–602. ISSN: 0967-2559. DOI: [10.1080/09672559.2020.1802971](https://doi.org/10.1080/09672559.2020.1802971).
- Klenk, Michael (2022). “Manipulation, Injustice, and Technology”. In: *The Philosophy of Online Manipulation*. New York: Routledge, pp. 108–131.
- Lahno, Bernd (2001). “Institutional Trust: A Less Demanding Form of Trust?” In: *Revista Latinoamericana de Estudios Avanzados* 15, pp. 19–58.

- Lehrer, Keith (1990). *Theory of Knowledge*. Boulder: Westview Press.
- Leibniz, Gottfried Wilhelm (2006). *Nouveaux Essais*. Ed. by Leibniz-Forschungsstelle Univ. Münster. Akademie Verlag, (Original work published 1765). ISBN: 9783050060750.
- Leslie, David (2019). *Understanding Artificial Intelligence Ethics and Safety*. URL: <http://arxiv.org/abs/1906.05684>.
- Levi, Isaac (1980). *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Cambridge: The MIT press.
- Leviathan, Yaniv and Yossi Matias (2018). *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*. URL: <https://research.google/blog/google-duplex-an-ai-system-for-accomplishing-real-world-tasks-over-the-phone/>.
- Lewis, Clarence Irving (1946). *An Analysis of Knowledge and Valuation*. LaSalle: Open Court.
- Lewis, David (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- (1994). “Humean Supervenience Debugged”. In: *Mind* 103.412, pp. 473–490. ISSN: 00264423, 14602113. URL: <http://www.jstor.org/stable/2254396>.
- (1997). “Finkish dispositions”. In: *The Philosophical Quarterly* 47.187, pp. 143–158. ISSN: 0031-8094.
- Lin, Xiaolin, Xuequn Wang, and Nick Hajli (2019). “Building E-Commerce Satisfaction and Boosting Sales: The Role of Social Commerce Trust and Its Antecedents”. In: *International Journal of Electronic Commerce* 23.3, pp. 328–363. DOI: [10.1080/10864415.2019.1619907](https://doi.org/10.1080/10864415.2019.1619907).
- Liske, Michael-Thomas (2009). “Harmonizing Modern Physics with Aristotelian Metaphysics. Leibniz’s Theory of Force”. In: *Issues in Metaphysics, Epistemology and Philosophy of Mind*. Berlin: De Gruyter, pp. 99–126.
- List, Christian and Philip Pettit (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
- Locke, John (1812). *The Works of John Locke*. Vol. I-X. Printed for W. Otridge and Son, [and 17 Others].
- Luhmann, Niklas (1980). *Trust and Power*. New York: Wiley.
- Martin, Charles B (1994). “Dispositions and Conditionals”. In: *The Philosophical Quarterly* 44.174, pp. 1–8. DOI: [10.2307/2220143](https://doi.org/10.2307/2220143).
- Mathiesen, Kay (2006). “The Epistemic Features of Group Belief”. In: *Episteme* 2, pp. 161–175. DOI: [10.3366/epi.2005.2.3.161](https://doi.org/10.3366/epi.2005.2.3.161).
- Matthias, Andreas (2003). “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata”. In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: [10.1007/s10676-004-3422-1](https://doi.org/10.1007/s10676-004-3422-1).
- McKenna, Michael (2012). *Conversation and Responsibility*. Oxford: Oxford University Press.
- McLeod, Caroline (2021). “Trust”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023 Edition. URL: <https://plato.stanford.edu/archives/fall2021/entries/trust/>.
- Medina, José (2020). “Trust and Epistemic Injustice”. In: *The Routledge Handbook of Trust and Philosophy*. Ed. by Judith Simon. New York: Routledge, pp. 52–63.

- Mele, Alfred R (1992). *Springs of Action: Understanding Intentional Behavior*. Cambridge: Oxford University Press.
- Miller, Tim (2017). “Explanation in Artificial Intelligence: Insights from the Social Sciences”. In: *Artificial Intelligence* 267, pp. 1–38. DOI: [10.1016/j.artint.2018.07.007](https://doi.org/10.1016/j.artint.2018.07.007).
- Mills, Eugene (1998). “The Unity of Justification”. In: *Philosophical and Phenomenological Research*, pp. 27–50. DOI: [10.2307/2653629](https://doi.org/10.2307/2653629).
- Mittelstadt, Brent Daniel et al. (2016). “The Ethics of Algorithms: Mapping the Debate”. In: *Big Data & Society* 3.2, pp. 1–21. DOI: [10.1177/2053951716679679](https://doi.org/10.1177/2053951716679679).
- Mumford, Stephen (1998). “Dispositions”. In: *Routledge Encyclopedia of Philosophy*. Taylor and Francis. DOI: [10.4324/9780415249126-N116-1](https://doi.org/10.4324/9780415249126-N116-1). URL: <https://www.rep.routledge.com/articles/thematic/dispositions/v-1>.
- Nguyen, C Thi (2019). “Trust as an Unquestioning Attitude”. In: *Oxford Studies in Epistemology*, Forthcoming. URL: <https://philpapers.org/archive/NGUTAA.pdf>.
- (2021). “Transparency is Surveillance”. In: *Philosophy and Phenomenological Research* Forthcoming Issue, pp. 1–31. DOI: [10.1111/phpr.12823](https://doi.org/10.1111/phpr.12823).
- Nickel, Philip J, Maarten Franssen, and Peter Kroes (2010). “Can We Make Sense of the Notion of Trustworthy Technology?” In: *Knowledge, Technology & Policy* 23.3, pp. 429–444. DOI: [10.1007/s12130-010-9124-6](https://doi.org/10.1007/s12130-010-9124-6).
- OECD (2017). “OECD Guidelines on Measuring Trust”. In: *OECD Publishing*. DOI: [10.1787/9789264278219-en](https://doi.org/10.1787/9789264278219-en).
- Olsson, Erik (2021). “Coherentist Theories of Epistemic Justification”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward Zalta and Uri Nodelman. Winter 2023 Edition. URL: <https://plato.stanford.edu/archives/win2023/entries/justep-coherence/>.
- OpenAI (2023). “GPT-4 Technical Report”. In: URL: <https://arxiv.org/abs/2303.08774>.
- Ortega, Alfonso et al. (2021). “Symbolic AI for XAI: Evaluating LFIT Inductive Programming for Fair and Explainable Automatic Recruitment”. In: *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*.
- O’Brien, Lilian (2022). “Intentional Agency”. In: *The Routledge Handbook of Philosophy of Agency*. New York: Routledge, pp. 109–117.
- O’Neil, Collin (2012). “Lying, Trust, and Gratitude”. In: *Philosophy & Public Affairs* 40.4, pp. 301–333. DOI: [10.1111/papa.12003](https://doi.org/10.1111/papa.12003).
- O’Neill, Onora (2002). *A Question of Trust*. Cambridge: Cambridge University Press.
- (2018). “Linking Trust to Trustworthiness”. In: *International Journal of Philosophical Studies* 26.2, pp. 293–300. DOI: [10.1080/09672559.2018.1454637](https://doi.org/10.1080/09672559.2018.1454637).
- Páez, Andrés (2019). “The Pragmatic Turn in Explainable Artificial Intelligence (XAI)”. In: *Minds and Machines* 29.3, pp. 441–459. DOI: [10.1007/s11023-019-09502-w](https://doi.org/10.1007/s11023-019-09502-w).
- Perloff, Richard M (2003). *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. New York: Lawrence Erlbaum Associates Publishers.
- Persson, Ingmar (2021). “Moral Status and Moral Significance”. In: *Rethinking Moral Status*. Ed. by Steve Clarke, Hazem Zohny, and Julian Savulescu. Oxford, pp. 108–122.

- Pettit, Philip (1995). “The Cunning of Trust”. In: *Philosophy & Public Affairs* 24.3, pp. 202–225. DOI: [10.1111/j.1088-4963.1995.tb00029.x](https://doi.org/10.1111/j.1088-4963.1995.tb00029.x).
- Placani, Adriana (2024). “Anthropomorphism in AI: Hype and Fallacy”. In: *AI and Ethics*, pp. 691–698. DOI: [10.1007/s43681-024-00419-4](https://doi.org/10.1007/s43681-024-00419-4).
- Pozzi, Giorgia (2023). “Testimonial Injustice in Medical Machine Learning”. In: *Journal of Medical Ethics* 49.8, pp. 536–540. DOI: [10.1136/jme-2022-108630](https://doi.org/10.1136/jme-2022-108630).
- Quine, Willard Van Orman (1960). *Word and Object*. Cambridge: The MIT Press.
- Reinhardt, Karoline (2022). “Trust and Trustworthiness in AI Ethics”. In: *AI and Ethics* 3.3, pp. 735–744. DOI: [10.1007/s43681-022-00200-5](https://doi.org/10.1007/s43681-022-00200-5).
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “Why Should I Trust You? Explaining the Predictions of Any Classifier”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17, pp. 1135–1144. DOI: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778).
- Ribera, Mireia and Agata Lapedriza (2019). “Can We Do Better Explanations? A Proposal of User-Centered Explainable AI.” In: *IUI Workshops* 2327, pp. 38–45. URL: <http://hdl.handle.net/10609/99643>.
- Rieder, Gernot, Judith Simon, and Pak-Hang Wong (2021). “Mapping the Stony Road toward Trustworthy AI”. In: *Machines We Trust: Perspectives on Dependable AI*. The MIT Press, pp. 27–39. DOI: [10.2139/ssrn.3717451](https://doi.org/10.2139/ssrn.3717451).
- Robbins, Scott (2019). “A Misdirected Principle with a Catch: Explicability for AI Keywords Ethics of AI”. In: *Minds and Machines* 29, pp. 495–514. DOI: [10.1007/s11023-019-09509-3](https://doi.org/10.1007/s11023-019-09509-3).
- Rougier, Nicolas P., Michael Droettboom, and Philip E. Bourne (2014). “Ten Simple Rules for Better Figures”. In: *PLoS Computational Biology* 10.9, pp. 1–7. ISSN: 15537358. DOI: [10.1371/JOURNAL.PCBI.1003833](https://doi.org/10.1371/JOURNAL.PCBI.1003833).
- Rousseau, Jean Jacques (1755). *Discourse on the Origin of Inequality*. New York: Dover Publications.
- Ruben, David Hillel (1990). *Explaining Explanation*. Ed. by Ted Honderic. London and New York: Routledge.
- Russell, Stuart J. and Peter Norvig (2009). *Artificial Intelligence. A Modern Approach*. Noida: Pearson Education Inc.
- Ryan, Mark (2020). “In AI We Trust: Ethics, Artificial Intelligence, and Reliability”. In: *Science and Engineering Ethics* 26, pp. 2749–2767. DOI: [10.1007/s11948-020-00228-y](https://doi.org/10.1007/s11948-020-00228-y).
- Ryle, Gilbert (1949). *The Concept of Mind*. London: Penguin.
- Sahoh, Bukhoree and Anant Choksuriwong (2023). “The Role of Explainable Artificial Intelligence in High-Stakes Decision-Making Systems: a Systematic Review”. In: *Journal of Ambient Intelligence and Humanized Computing* 14.6, pp. 7827–7843. DOI: [10.1007/s12652-023-04594-w](https://doi.org/10.1007/s12652-023-04594-w).
- Samek, Wojciech et al. (2016). “Evaluating the Visualization of What a Deep Neural Network Has Learned”. In: *IEEE Transactions on Neural Networks and Learning Systems* 28.11, pp. 2660–2673. DOI: [10.1109/TNNLS.2016.2599820](https://doi.org/10.1109/TNNLS.2016.2599820).
- Sartorio, Carolina (2016). *Causation and Free Will*. Oxford: Oxford University Press.
- Saultz, John (2020). “Failed Trust”. In: *Fam Med*. 52.9, pp. 621–622. DOI: [10.22454/FamMed.2020.818694](https://doi.org/10.22454/FamMed.2020.818694).

- Shin, Donghee (2021). “The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI”. In: *International Journal of Human Computer Studies* 146.102551, pp. 1–10. DOI: [10.1016/j.ijhcs.2020.102551](https://doi.org/10.1016/j.ijhcs.2020.102551).
- Simon, Judith (2010). *Knowing Together: A Social Epistemology for Socio-Technical Epistemic Systems*. Dissertation.
- ed. (2020). *The Routledge Handbook of Trust and Philosophy*. New York: Routledge.
- Smith, Adam (1759). *The Theory of Moral Sentiments*. Oxford: Clarendon Press.
- Sparrow, Robert (2012). “Killer Robots”. In: *Journal of Applied Philosophy* 24.1, pp. 62–77.
- Spohn, Wolfgang (1988). “Ordinal Conditional Functions: A Dynamic Theory of Epistemic States”. In: *Causation in Decision, Belief Change, and Statistics*. Irvine: Springer, pp. 105–134.
- (2018). “Epistemic Justification: its Subjective and its Objective Ways”. In: *Synthese* 195.9, pp. 3837–3856. DOI: [10.1007/s11229-017-1393-0](https://doi.org/10.1007/s11229-017-1393-0).
- Strawson, Peter Frederik (1962). “Freedom and Resentment”. In: *Proceedings of the British Academy* XLVIII, pp. 1–25. DOI: [doi:10.1017/9780691200972-010](https://doi.org/10.1017/9780691200972-010).
- Suchman, Lucy (2006). *Human-Machine Reconfigurations*. Cambridge: Cambridge University Press.
- Talbert, Matthew (2024). “Moral Responsibility”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024 Edition. URL: <https://plato.stanford.edu/archives/sum2024/entries/moral-responsibility/>.
- Tomsett, Richard et al. (2018). “Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems”. In: *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. URL: <https://orca.cardiff.ac.uk/id/eprint/112597/>.
- United Nations (2021). “Recommendation on the Ethics of Artificial Intelligence”. In: URL: <https://unesdoc.unesco.org/ark:/48223/pf0000379920>.
- Van House, Nancy (2003). “Science and Technology Studies and Information Studies”. In: *Annual Review of Information Science and Technology (ARIST)* 38, pp. 3–86. ISSN: 0066-4200.
- Vetter, Barbara (2013). “Multi-Track Dispositions”. In: *The Philosophical Quarterly* 63.251, pp. 330–352. DOI: [10.1111/1467-9213.12024](https://doi.org/10.1111/1467-9213.12024).
- (2014). “Dispositions without Conditionals”. In: *Mind* 123.489, pp. 129–156. DOI: [10.1093/mind/fzu032](https://doi.org/10.1093/mind/fzu032).
- Vredenburgh, Kate (2022). “Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace”. In: *Canadian Journal of Philosophy* 52.1, pp. 78–92. DOI: [10.1017/can.2021.39](https://doi.org/10.1017/can.2021.39).
- Wachter, S, B Mittelstadt, and C Russell (2018). “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: *Harvard Journal of Law and Technology* 31.2. ISSN: 0897-3393.
- Wang, Danding et al. (2019). “Designing Theory-Driven User-Centric Explainable AI”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. DOI: [10.1145/3290605.3300831](https://doi.org/10.1145/3290605.3300831).

- Watson, Gary (1996). “Two Faces of Responsibility”. In: *Philosophical Topics* 24.2, pp. 227–248. DOI: [10.5840/philtopics199624222](https://doi.org/10.5840/philtopics199624222).
- Woodward, James (2004). *Making Things Happen. A Theory of Causal Explanation*. New York: Oxford University Press.
- Xiong, Haoyi et al. (2024). “Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective”. In: *Preprint*. DOI: [10.48550/arXiv.2401.04374](https://doi.org/10.48550/arXiv.2401.04374).
- Yan, Zheng et al. (2013). “Exploring Trust of Mobile Applications Based on User Behaviors: an Empirical Study”. In: *Journal of Applied Social Psychology* 43.3, pp. 638–659. DOI: [10.1111/j.1559-1816.2013.01044.x](https://doi.org/10.1111/j.1559-1816.2013.01044.x).
- Zerilli, John et al. (2019). “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?” In: *Philosophy & Technology* 32.4, pp. 661–683. DOI: [10.1007/s13347-018-0330-6](https://doi.org/10.1007/s13347-018-0330-6).



