Trust, Testimony, and Transmission

Essays in Social Virtue Epistemology

Ísak Andri Ólafsson

Submitted in fulfilment of the requirements for the Degree of

Doctor of Philosophy

Philosophy

School of Humanities

College of Arts

University of Glasgow

December 2024

Dissertation Outline

This thesis consists of five distinct essays within social virtue epistemology, each of which can stand independently, yet all engage with fundamental ideas surrounding trust, testimony, and knowledge transmission. The first two chapters explore knowledge transmission and testimony through a virtue epistemological lens, emphasising the challenges of accounting for testimonial knowledge while maintaining a connection between knowledge and credit. I introduce types of knowledge transmission that do not rely on joint agency or shared intentions, challenging a prominent view in virtue epistemology. I present a type of a credit view that can defend one of the fundamental doctrines of credit views, that knowledge always entails credit, from challenging counterexamples. Trust and testimony both facilitate connections between individuals, making them central to our understanding of how knowledge is shared in social contexts. The third chapter aims to further our understanding of the nature of trust by placing the spotlight on trust features that have gone largely unnoticed, namely, their temporal elements. By expanding on these features, we can make meaningful distinctions between instances of trust that have generally been considered interchangeable. These distinctions and related concepts highlight the subtle differences that meaningfully impact how we approach trust. In the fourth chapter, the focus shifts to epistemic groups in the context of gatekeeping. Epistemologists should be interested in trust, testimony, and transmission as they relate to individuals, but groups are an interesting epistemic subject in their own right. This chapter examines the distinct epistemic roles groups play in shaping the beliefs of their members, and how individuals can benefit from being part of a collective. I then present conditions for justified epistemic gatekeeping and consider what kinds of groups are most capable of fulfilling those conditions. In the last chapter, I consider how to define general artificial intelligence. It is difficult to place large language models within epistemology. At times, they act like epistemic agents, seemingly capable of producing and transmitting knowledge, yet they often appear incompetent and incapable of performing simple tasks. I propose a virtue-theoretic distinction between narrow and general artificial intelligence, in the hopes that it can contribute to our understanding of what makes AI trustworthy, and whether we should think of their predictions as knowledge.

Table of Contents

Dedication

I dedicate this thesis to the memory of Elsa H. Sigurðardóttir, who always encouraged my curiosity.

Acknowledgements

I would like to thank my supervisory team, Adam Carter, Christoph Kelp, and Mona Simion, for their guidance, support, and encouragement throughout this process. This dissertation would not have been possible without their contributions.

I extend my thanks to the staff at the University of Glasgow and to the friends I made during my time there. Special thanks to Cian Brennan, Rory Aird, and Ross Patrizio for reading my papers.

I wish to express my gratitude to Finnur Dellsén, Elmar Unnsteinsson, and Jón Ólafsson for their contributions to my academic development over the years.

This dissertation was written as part of the Leverhulme-funded *A Virtue Epistemology of Trust* project. I am grateful to the Leverhulme Trust for supporting my research and to the Royal Institute of Philosophy for their additional support.

Some of the work in this thesis has benefitted from feedback provided by anonymous referees for *Synthese*.

I am deeply grateful to my family for their unconditional love and encouragement, and to my friends for their endless patience.

Finally, I want to thank Natalía Lind Jóhannsdóttir for everything.

Author's Declaration

I declare that this thesis is the result of my own work except where explicit reference is made to the contribution of others. This work has not been submitted for any other degree at the University of Glasgow or any other institution. This work has received generous financial support from the Leverhulme Trust and the Royal Institute of Philosophy.

Ísak Andri Ólafsson

Published Material

At the time of the submission of this thesis (December 2024), chapter 1 has been published:

Ólafsson, Ísak Andri (2023). Unwanted Knowledge Transmission. *Synthese* 201, 162:1-17

Introduction

We can learn from others. Theories on trust, testimony, and knowledge transmission all relate to this fact in some way, and these social aspects of knowledge can contribute to our understanding of how we come to know from others. This thesis consists of five distinct essays in social virtue epistemology that concern these three pillars of epistemic communication: trust, testimony, and transmission.

The chapters all relate to these fundamental concepts, but they can stand on their own, providing an opportunity to demonstrate the many facets and applications of social virtue epistemological ideas. The aim of each chapter is to introduce novel ideas that expand the social dimensions of knowledge, and more generally identify problems and solutions within social epistemology that interface with virtue epistemology under these themes.

The aim of the first chapter is to show that it is possible to transmit knowledge without the speaker and hearer sharing an intention to do so. I introduce examples of unwanted knowledge transmissions, which come in at least four types: irrational, unintended, resented, and achievement-depriving. These types of transmissions place transmission theories that rely on joint agency and shared intentions in a dilemma. Namely, that competent joint agency is either not present in many typical cases of testimonial knowledge exchanges, which makes those cases susceptible to Lackey's dilemma, or it is present in a diminished form that cannot produce the sort of shared credit that is needed to defend against Lackey's dilemma.

In the second chapter I argue that knowledge always entails credit. I begin by motivating why we should concern ourselves with virtue-theoretic credit views of knowledge and assess a standard individual version of a credit view. I then introduce a challenging counterexample that shows how such views are susceptible to Lackey's dilemma. In an attempt to escape Lackey's dilemma, I present a social species of a credit view, inspired by speaker's accounts of knowledge transmission. Unfortunately, the social credit view runs into other counterexamples involving a speaker that transmits knowledge they themselves do not possess.

In a final attempt, I construct the basic credit view. On this view, having cognitive contact with reality is seen as a cognitive success that follows an achievement structure. I argue that if there is testimonial knowledge, then there has to have been a basic epistemic achievement somewhere in the testimonial chain, where a basic epistemic achievement is lesser than knowledge, but is nevertheless a genuine epistemic achievement, and creditable as such. In non-testimonial cases I rely on the stronger individual credit view. This allows for an account that can keep a strong version of a credit view in non-testimonial cases, while endorsing a weaker version of credit in cases of testimony. The counterexamples aimed at the social credit view do not pose a challenge to a fundamental claim of the extended family of credit views; that knowledge always entails credit. Furthermore, I identified that some paradigmatic counterexamples to transmission theories more generally are all structured in a way that utilises two kinds of unreliability of epistemic agents. After systematically reviewing the kinds of cases that can occur, I presented the most influential ones and responded to them.

In the third chapter I argue that there are some underdeveloped features of trust that can provide a meaningful distinction between certain instances of trust that have often been seen as being interchangeable. I make a distinction between trust affirmations and ongoing trusting relations, which are again separated into definite ongoing trust and indefinite ongoing trust.

I further introduce the concepts of trust resolutions, confirmation monitoring, disclosed monitoring, and undisclosed monitoring, with the aim of clarifying the nature of trust. These distinctions are then employed in an attempt to challenge the notion that monitoring always undermines trust, argue that therapeutic trust requires a kind of monitoring to be maximally successful, and how these temporal elements affect precautionary measures and preemptive reasons for trusting, particularly as they relate to risks of betrayal and disutility. I then use the terms introduced in the chapter to challenge and explore three different trust accounts, namely, Emma Gordon's idea that monitoring can facilitate trust, Arnon Keren's claim that one cannot trust without responding to preemptive reasons, and Wanderer and Townsend's attempt to reconcile trust and rationality.

The fourth chapter explores epistemic gatekeeping in groups. I argue that gatekeeping, as a way to facilitate cheap knowledge transmissions within groups by relying on distribution norms

inside the group while falling back on acquisition norms when assessing testimony from outside the group, faces two problems that are intrinsic and unescapable.

The first problem occurs if a member of a gatekeeping group acquires a false belief after following (generally) reliable acquisition norms. If they then testify their false information to someone else within the group, the hearer is in a vulnerable position because they will be relying on distribution norms when assessing testimony that originates from within the group. The risk is that everyone in the group will eventually form the respective false beliefs because of their reliance on distribution norms which are more lenient than the stricter acquisition norms.

The second problem occurs when a member of a group hears good testimony, but erroneously dismisses it as being false. This can lead to knowledge deprivation, or insufficient coverage, as the members of the group that might otherwise form the relevant true belief are unable to do so as a result of trusting in the group's gatekeeping process.

I then show that there is a correlation between the epistemic labour the group invests in the gatekeeping process, and the risk of the two problems. As groups increase their epistemic work by shifting further towards acquisition norms, the chance of the gatekeeping problems materialising decreases. However, for groups to fully eliminate the two gatekeeping problems, they have to fully rely on acquisition norms, which means that knowledge transmissions are as expensive as they were without gatekeeping. If this is the case, then it will be difficult for groups to justify gatekeeping.

Finally I propose three conditions of justified gatekeeping that minimise the risks associated with gatekeeping by reducing the severity of the gatekeeping problems rather than their frequency. The conditions are, in short, that groups must be a part of an epistemically interrelated community of groups, they must have sufficient epistemic autonomy, and they must be subjectively epistemically rational. We then suggest that scientific communities are structured in a way that makes expert groups ideal candidates for gatekeeping in a justified manner.

In the fifth and last chapter, I present a virtue-theoretic distinction between general AI (GAI) and narrow AI (NAI) systems, relying on Sosa's virtue-theoretic framework. I begin by making a distinction between narrow AI and general AI. Then, I note that competences, under the virtue-theoretic framework being used in the chapter, are understood as special kinds of dispositions of an agent to perform well in a given domain. I then examine which domains are relevant to AI. Additionally, I consider what constitutes an AI attempt, and what kinds of actions they are attempting. I then demonstrate that the difference between general AI and narrow AI is rooted in the range of domains the AI is proficient in.

I argue that for an AI system to become a general AI, it needs to have the constitutional competence in epistemic domains to reflect on its complete first-order competence. In contrast, NAI have complete competence in some non-epistemic domain. This suggests that the difference between NAI and GAI is not simply a matter of competence within a specific domain. To put this differently, NAI cannot become a GAI just by improving their competence in their respective domains unless they advance in the epistemic domains. I then argue that GAI could potentially reach the lowest creditability for knowledge as the predictions made by them will be aptly apt in a way that avoids the limitations NAI faces. I finally argue that GAI, if it existed, would potentially have the competence to predict in a manner that constitutes reflective knowledge (global competence).

At the end of the five chapters, I write a short conclusion that summarises my findings in each chapter. I then suggest a potential area of further exploration that relates to indefinite and definite trust, and how some situations can cause trust to behave in a different way than we would expect.

Chapter 1: Unwanted knowledge Transmission

Abstract. John Greco (2020) sets out to create a unified virtue-theoretic account of knowledge generation and transmission in his book *The Transmission of Knowledge*. One of the advantages of his view is that it can defend the achievement view from a known strand of counterexamples. To accomplish that, he relies on joint agency being essential to knowledge transmission. Joint agency can be characterised as a sort of interdependent and interactive cooperation between speaker and hearer that share an intention to transmit knowledge. This chapter introduces the phenomenon of unwanted knowledge transmission which shows that joint agency, as Greco presents it, is not essential to knowledge transmission. Four types of unwanted knowledge transmissions will be introduced that pose threats to Greco's account by showing that shared intention, a key characteristic of joint agency, is not present in all instances of knowledge transmission. Finally, some potential ways to defend Greco's view will be considered and discussed.

Introduction

John Greco (2020), in his book *The Transmission of Knowledge,* argues for an anti-reductionist theory of knowledge transmission that sees knowledge transmission and knowledge generation as two distinct phenomena, where one cannot be reduced to the other and each requires a distinct approach. So, in Greco's view, there are two ways of coming to know.

Firstly, individuals can generate knowledge from their competent agency (Greco 2020, 98). This competent agency is manifested by the individual when they competently use their abilities, such as their perception, to come to know something for themselves. As an example, imagine that you are standing in a field and a sheep walks up to you. Using your reliable faculties, you come to know that there is in fact a sheep standing before you based on your visual perception.

Secondly, individuals can receive transmitted knowledge by being a part of a competent joint agency (where both hearer and speaker are acting competently), where the true belief

acquisition is attributable to the competent joint agency itself rather than the competent participation of the participating individuals (Greco 2020, 98-99).

So, knowledge generation is when one comes to know of their own accord, while knowledge transmission is when one comes to know from someone else without the usual epistemic burden that knowledge generation generally entails (Greco 2020, 1). Epistemic burden here can be thought of as the work required to generate knowledge by one's own efforts using one's own cognitive and sensory abilities. Knowledge transmission, then, allows for a division of epistemic labour[1] in a way that knowledge generation does not (Greco 2020, 3-4).

According to Greco, knowledge transmission essentially involves *telling*, a speech act similar to assertion but with the clear purpose of sharing knowledge[2], and it is a kind of joint agency[3], in which individual agents act *together*. Joint agency can be characterised as "a network of shared intentions and common understanding between the participating actors, as well as specific kinds of interdependence" (Greco 2020, 47). Furthermore, "trust is essential to joint agency in general, and joint agency is essential to knowledge transmission. It follows that trust is essential to one important way of coming to know" (Greco 2020, 48).

In cases of knowledge, true belief is always attributable in some way to the competent cognitive agency of the knower (Greco 2020, 99). Joint agency and joint achievement explain how the success can be credited to the hearer's competent agency even if that agency is not the most salient part of the success. When a success is directly attributable to the joint agency itself, the success is also indirectly attributable to the individuals that are a part of the joint agency to the extent they contributed to the joint agency in a competent way (Greco 2020, 99).

A motivating factor for Greco here is to avoid a known strand of counterexamples to traditional virtue epistemological views, that conceive knowledge as an individual achievement of the knower, by showing how they fail to account for the inherent social dimensions of knowledge

[1] This is referring to Sandy Goldberg's (2011) thesis of the division of epistemic labour.
[2] Christoph Kelp and Mona Simion (2022) have investigated the relation between knowledge sharing and assertion in their book *Sharing Knowledge.* In their view, the function of assertion is to share knowledge with others.
[3] Abraham Sesshu Roth (2017) has written a helpful summary of shared agency.

(Greco 2020, 94). One prominent counterexample of this sort is Lackey's creditworthiness dilemma.[4]

Greco's account, if successful, can defend the achievement view from such counterexamples by stating that in cases of knowledge transmission, the hearer depends on the speaker as a cooperating agent in joint activity and the hearer's true belief is thus attributable to the competent joint agency of both speaker and hearer (Greco 2020, 101). This preserves the advantages of the traditional achievement view, namely, that knowledge can still be considered intrinsically and finally valuable because it is an achievement, and knowledge is more valuable than mere true belief because we value achievements over mere lucky successes (Greco 2020, 101). Greco's view just adds that knowledge is sometimes a joint achievement that is attributable to competent joint agency.

Joint agency thus plays a significant role in Greco's unified virtue-theoretic account of knowledge generation and transmission, an account he finds to be a "significant revision to traditional virtue epistemology" (Greco 2020, 98).

Here is the plan. The first section will elaborate on Greco's view on knowledge transmissions and how it relates to joint agency. In the next section we will examine joint agency in more

[4] For an understanding of Jennifer Lackey's creditworthiness dilemma, consider her case of the Chicago visitor: "Having just arrived at the train station in Chicago, Morris wishes to obtain directions to the Sears Tower. He looks around, approaches the first adult passerby that he sees, and asks how to get to his desired destination. The passerby, who happens to be a lifelong resident of Chicago and knows the city extraordinarily well, provides Morris with impeccable directions to the Sears Tower by telling him that it is located two blocks east of the train station. Morris unhesitatingly forms the corresponding true belief." (Lackey 2009, 29). Morris seems to form a true belief because of the testimony of the speaker. This means that his competent agency is not a salient enough reason for his true belief, which goes against the achievement view of virtue epistemology. This is problematic because if we allow Morris to rely on the speaker's abilities, as well as his own, then Gettier-type situations cannot be ruled out. To see why that is, consider a Gettier-type case where the environment is less suited for the formation of true beliefs such as a fake barn case, in which an epistemic agent sees a barn in the distance and forms the corresponding true belief. However, the barn in question is the only real barn among many barn façades. In this example, the agent's abilities and efforts are an important factor of how they came to know that there is a barn in the distance, but they are still lucky that their belief is true (Lackey 2009, 33). Morris is in a relevantly analogous situation and thus vulnerable to these types of Gettier-situations; he was lucky that he asked a passerby that happened to know the city extraordinarily well. To formulate Lackey's dilemma in a succinct way, consider that in cases of knowledge, S's competent agency explains why S has a true belief and if S's competent agency must be a salient enough reason for S's true belief, then many cases of testimonial knowledge are ruled out (such as Chicago visitor). However, if S's competent agency does not need to be the most salient reason for S's true belief, instead being just one important reason for it, then Gettier-type cases are allowed to go through. So, we either rule out cases like Chicago visitor or we allow Gettier-type cases, in either case the achievement view is in trouble.

detail and highlight the necessary conditions for it. We will introduce the epistemological phenomenon of *unwanted knowledge transmission* (UKT) that is at odds with the essential characteristics of joint agency, and thus with Greco's account of knowledge transmission. For the time being, UKT can be defined as successful knowledge transmissions that are undesired by either speaker or hearer, e.g., when someone spoils the ending of a movie for you by telling you the plot twist.

Note that UKT cases are not meant to refute the claim that trust is essential to joint agency, but that joint agency, regardless of how it relates to trust, is not essential to knowledge transmission. If any of these cases successfully show that joint agency is not essential to knowledge transmissions, then Greco's view is compromised. We will then consider potential objections to the problems UKT introduce.

Shared Intention and Knowledge Transmission

The speech act of *telling*[5] and the corresponding joint action are closely related. When knowledge transmission is successful, the speaker tells the hearer something, manifesting their intention to share the knowledge with the hearer. Furthermore, the hearer understands and shares this intention, thus becoming a participant in the joint action of sharing knowledge (Greco 2020, 49). For example, say you are visiting a friend that lives in a city you are unfamiliar with. You want to know where the nearest botanical garden is, so you ask your friend and they tell you where it is. In this example, you and your friend act together to ensure that the knowledge transmission goes smoothly. You ask your friend in a clear manner and they try to convey their knowledge to you in a competent manner. You are aware of the general content of what they are about to tell you and you anticipate their reply to help you on your way to the garden.

According to Greco, speech acts and other kinds of actions require that the world cooperates when trying to do something for it to achieve its aim, that is, trying to do X is not sufficient for X to happen; there must be cooperation involved (Greco 2020, 49). The act of telling, as a

[5] Philosophers who have focused on the speech act of telling in instances of testimonial knowledge are, for example, Edward S. Hinchman (2005), Richard Moran (2005), (2018), and Charlie Pelling (2014).

species of assertion, follows general characteristics[6] of speech acts, and "is a social institution for the transmission of knowledge" (Greco 2020, 49-50).[7]

The speech act of telling is the action involved in testimonial exchanges and it expresses a kind of interdependency between speaker and hearer, which is also an essential component of joint agency. Joint agency, as it appears in testimonial exchanges, can be "characterised by a special sort of cooperation between speaker and hearer" (Greco 2020, 18) and it essentially involves trust relations between them (Greco 2020, 18).

In any communicative exchange, the speaker intends to be understood by their hearer and the hearer understands that this is the intention of the speaker. Furthermore, the speaker and the hearer cooperate to achieve the intended result and they depend on each other to make the appropriate contributions to the exchange (Greco 2020, 57). Greco emphasises that joint agency is more demanding than general communicative exchanges because knowledge transmissions involve a shared intention to impart relevant information (Greco 2020, 57).

Think of two people walking to the same restaurant, but neither person knows where the other is headed (Greco 2020, 54). In this scenario we see that there is no shared intention between them even though they both have an intention to walk to the restaurant, but if they were walking to the restaurant together[8] then they would share their intention of going to the restaurant, and they would both be aware that this intention is present (Greco 2020, 54). This kind of teamwork[9], that manifests as a shared intention, is a hallmark of joint agency[10] and

[6] Hinchman (2005, 567) has defined telling along these lines: S tells A that *p* iff A recognises that S, in asserting that *p*, intends: that A gain access to an epistemic reason to believe that *p*, that A recognise S's (2)-intention, and that A gain access to the epistemic reason to believe that *p* as a direct result of A's recognition of S's (2)-intention (Leonard 2021).

[7] Greco adopts this from Elizabeth Fricker's view on telling, where she states that "telling is a social institution for the spreading of knowledge" (Fricker 2006, 594-596).

[8] To learn more about these kinds of cases, shared activity, and their characteristic features, see Michael E. Bratman (1992). For further discussion on walking together, see Margaret Gilbert (1990).

[9] For a discussion on group competence and teamwork, see Ernest Sosa (2007, 94-95).

[10] Other characteristics of joint agency that are generally accepted in the literature are as follows; (1) it involves a shared intention between the joint agents, that is, they understand their action as something that they are doing together, (2) it involves planning on the part of the joint agents about how they are going to carry that action out together, (3) it is interactive, so the actions of one agent affect the actions of the other, (4) it is interdependent, so an individual agent cannot perform the joint action alone in the same way; each agent is doing a part of the joint action (Greco 2020, 55-56).

"[k]nowledge transmission is to be understood in terms of success due to the competent joint agency of speaker and hearer acting together" (Greco 2020, 19). As shared intention is an integral part of joint agency, we can see that to effectively argue against Greco's view it is enough to show that there can be knowledge transmission without shared intention.[11] We can now present Greco's account of knowledge transmission.

> **Greco's Account of Knowledge Transmission:** Knowledge that *p* is transmitted from a speaker S to a hearer H just in case S successfully tells H that *p*. And that happens just in case:
>
> (1) S knows that *p*.
> (2) S asserts that *p* with the intention of sharing knowledge that *p* with H.
> (3) H understands and shares S's intention.[12]
> (4) S and H act jointly so as to bring about their shared intention (i.e., so as to consummate the speech act in condition 2) (Greco 2020, 57).

Greco's goal is to create a new unified virtue-theoretic account of knowledge generation and transmission and he relies heavily on this account of knowledge transmission to do so. We can

[11] Many thanks to an anonymous referee whose comments prompted this clarification.

[12] Some clarification is needed here to explain how the hearer can share the speaker's intention to tell the hearer *that p* when the hearer does not know the particular proposition the speaker intends to transmit. One way Greco could explain this is by saying that we, as epistemic agents, have a standing disposition of intention as communicators. Unfortunately, Greco does not say anything about the temporal duration of the intention. This is not problematic in cases where the speaker is responding to a specific inquiry (for example, you ask your friend if they are coming to your party and they give you an answer) (Greco 2020, Case 4, p. 33), but it is difficult to accept that third-grade students shared their teacher's intention to tell them that France is in Europe (Greco 2020, Case 5, p. 33). Note that Greco addresses how a hearer can share a speaker's intention even when the hearer is not the one doing the telling, but he does not explain how they can share the specific intention of sharing knowledge *that p*. One attempt to address this is to reformulate Greco's account of knowledge transmission to have speakers and hearers intend to share knowledge in general, instead of some particular knowledge *that p*. However, as will be shown in the objections section of this chapter, going this route makes it difficult for Greco's view to explain what makes knowledge transmission special compared to e.g. communicative exchanges. To clarify, imagine two people that only share the intention to walk to any restaurant together. Eventually they arrive at a restaurant that only one of them (or, in some cases, neither of them) wanted to go to. This would still be considered an achievement that could be credited to their joint agency just as if they had arrived at a restaurant that they both liked, which seems undesirable to me. This reformulated account is also susceptible to Lackey's dilemma as it relies on this weaker notion of shared intention to be the most salient reason for how a hearer comes to know a proposition from testimony. I am grateful to an anonymous referee for prompting further clarification on this point.

derive a core thesis of joint activity as it relates to knowledge transmission using conditions (3) and (4) of Greco's account. Call this Greco's Shared Intention Principle.

> **Greco's Shared Intention Principle**: A speaker transmits knowledge that *p* to the hearer only if the speaker intends to inform the hearer that *p*, the hearer understands and shares this intention, and the two successfully cooperate to execute their shared intention successfully.

UKT cases challenge Greco's account of knowledge transmission as they exhibit successful knowledge transmissions that do not follow the shared intention principle (conditions (3) and (4) of Greco's account of knowledge transmission), and thus cannot be representative of joint agency. We can spot a potential objection straight away that goes like this: The features of joint agency that UKT attacks might be characteristic of joint agency, but they are not essential to it. Greco specifically talks about characteristic features of joint agency instead of essential features to avoid disputes about which features are necessary for the definition of joint agency, as he is arguing that transmission shares all of the characteristic features of joint agency and that stands regardless of whether the features of joint agency are characteristic or essential (Greco 2020, 55).

Even if this is the case, the joint agency features that Greco introduces[13] are the ones he uses to advance his view, and he argues that transmission shares all the characteristic features of joint agency regardless of whether they are essential or not. We see that even if Greco would be using a notion of joint agency that would lack features some would consider to be essential, then this less-than joint agency is still essential to his view, and the challenges that arise from UKT cases would remain the same. We can now introduce four distinct types of UKT that demonstrate how there can be knowledge transmission without shared intention or joint agency.

[13] Because they are characteristic and common of joint agency (Greco 2020, 55).

Unwanted Knowledge Transmission

UKT cases show that it is possible to transmit knowledge without shared intention[14] between speaker and hearer. The aim here is to demonstrate how Greco's account breaks down when facing this phenomenon.

We will present four types of cases that pose serious problems to Greco's shared intention principle. These cases all have in common that there is knowledge transmission[15] happening that is unwanted by either speaker or hearer (or both, in some cases). First, we will look at irrelevant knowledge transmissions, where the hearer is largely indifferent to the knowledge that is being transmitted. Then we will introduce unintended knowledge transmissions, where either hearer or speaker has no intention for knowledge transmission to take place. Next, we will look at resented knowledge transmissions, where the knowledge being transmitted is not only unintended but actively resented by the hearer. Finally, we will present achievement-depriving knowledge transmissions, which are resented because they can deprive someone of earning an achievement in a virtue-theoretical sense.

Irrelevant Knowledge Transmission

Irrelevant knowledge transmission is a type of UKT where a hearer comes to know something by transmission but is indifferent as to whether or not they came to know it. Irrelevant knowledge transmissions of this sort are prevalent in our day-to-day lives, but they are mostly harmless; they generally do not get in the way of us coming to know the things we want to know. Nevertheless, there is still something epistemically interesting about them; here is an example.

> MARVEL FANATIC**.** One of your friends just keeps talking about the Marvel universe. This friend goes on and on about superheroes, the multiverse, infinity stones, and so on. After many years of close friendship, you realise you have slowly come to know

[14] And without joint agency, as shared intention is one of the conditions of joint agency.
[15] It could be argued that some instances of unintended knowledge transmission are examples of knowledge generation rather than transmission.

> extraordinarily many things about this Marvel universe. However, you have never been interested in knowing any of those things; you don't even like superheroes.

We see that the hearer does not share the speaker's intention of transmitting knowledge in the same way they would in regular cases where both speaker and hearer share an intention to transmit knowledge.[16] This innocuous example hints that Greco's demand for shared intention might become problematic for his view; at the very least, here, the intention that the speaker's knowledge be transmitted is not shared by speaker and hearer to the same degree. But what if the speaker or the hearer (or both) did not have any intention of partaking in knowledge transmission? To answer that, let us turn our attention towards unintended knowledge transmissions to see how knowledge transmissions can occur without apparent intention.

Unintended Knowledge Transmission

Here are three kinds of knowledge transmissions that are unintended.

> EAVESDROPPING: The speaker does not intend for knowledge to be transmitted (but the hearer does).
>
> EAVESTALKING[17]: The hearer does not intend for knowledge to be transmitted (but the speaker does).
>
> OVERHEARING: Neither speaker nor hearer intend for knowledge to be transmitted.

For clarity, we will now present three corresponding examples that show how these distinct kinds of unintended transmissions appear to us. Consider this case of eavesdropping in which

[16] Greco employs a notion of intention that gives rise to a generality problem. We could argue that the hearer intends *de re* to have a conversation, in the sense that the hearer intends to have a specific non-UKT conversation with someone without intending *de dicto* to partake in conversations that could contain UKT. Still, the problem here is that the intention, as an operator, is in a referentially opaque context, where the identity conditions for intentions would need to be laid out before we can commit to a specific reading of Greco's notion of intention.

[17] The origin of the term eavesdropping comes from the obsolete noun eavesdrop, which is the ground on to which water drips from the eaves (and eavesdropping is thus when someone stands within the eavesdrop of a house, intending to listen in on a conversation inside the house). I named the corresponding speaker-intent scenario eavestalking, in which someone speaks within the eavesdrop of a house intending to be heard by the people inside the house.

a speaker transmits knowledge to a hearer but the speaker had no intention of transmitting knowledge.

> CHRISTMAS PRESENT**.** Parents are discussing their child's Christmas present. Unbeknownst to them, their curious child is eavesdropping on their conversation and comes to know what they will be getting for Christmas.

Eavesdropping cases are well known in the epistemic literature of assurance[18] and telling[19]. When someone eavesdrops on a conversation, they are not issued assurance because the speaker does not intend for the eavesdropping hearer to believe what they say, therefore failing to satisfy the conditions of telling as a speech act. Even though the intended speaker was issued assurance and the eavesdropping hearer was not, the assurance in question seems epistemically superfluous, that is, it does not seem to affect the epistemic status of the eavesdropping hearer's belief (Leonard 2021). Now, even if CHRISTMAS PRESENT could not be said to be a case of telling as a speech act, we can still consider other cases that seem to satisfy the conditions of telling. Look at the following case of eavestalking, in which a speaker transmits knowledge to a hearer, but the hearer had no intention of partaking in knowledge transmission.

> GIFT SUGGESTION. Drew is having some friends over for coffee while their spouse, Jordan, is working from home in the next room. Drew has their birthday coming up next week and is worried that Jordan doesn't know what birthday gift to get them. Drew proceeds to talk loudly about their birthday in the hope that Jordan might hear them. As it happens, Jordan can't help but hear what Drew said and comes to know that Drew would like a new coffee machine.

Finally, consider the following case of overhearing in which a speaker transmits knowledge to a hearer and both desire the knowledge transmission to take place, but neither intends for it to happen.

[18] See e.g., (Lackey 2008), (Leonard 2021), (Owens 2006), (Schmitt 2010, 216-242).
[19] Referring to telling as a speech act.

> DISTRACTED STUDENT. A philosophy professor wants to transmit knowledge to all their students. One student does not pay attention and does not come to know the things the professor said during a lecture. Afterwards, the student accidentally overhears the professor speaking to a colleague about the presentation they just delivered and comes to know the relevant material.

This is akin to two friends walking side by side to a restaurant without either friend realising that they are walking next to each other. It is strange to say that those friends are walking to the restaurant together in the same way as they would if they had planned to go together. In these three cases we see that there is no cooperation beyond some fundamental application of cognitive and sensory faculties. We also see that even though there is no competent joint activity taking place, and the requirements of Greco's shared intention principle are not met (as the speaker and hearer do not share their intention with each other), these cases still seem to portray testimonial knowledge exchanges. However, if the examples presented here are allowed as knowledge transmissions, then the problems that arise are significant and widespread, not just for Greco's view but numerous epistemological views that concern testimony and knowledge transmission.

One way to fight back against these unintended knowledge transmission cases is to say that they are not cases of knowledge transmission at all, but rather just typical examples of knowledge generation. Greco has said as much; communicative exchanges in general do not necessarily amount to knowledge transmission and testimony can sometimes generate knowledge instead of transmitting it (Greco 2020, 26). However, as we will see, it is challenging to explain why these commonplace testimonial exchanges should not be considered knowledge transmissions, and even if one could do so convincingly, then Greco's theory would be severely limited in scope. It would cease to be a theory of paradigmatic cases of testimonial knowledge exchanges and could only be applied to a seemingly artificially constructed subset of some such cases.

Furthermore, even if we were to concede this line of argument completely, there are still other kinds of knowledge transmissions that pose greater problems to Greco's shared intention principle and thus his account of knowledge transmission. To see where we are headed, consider

the following case of overhearing that shows a knowledge transmission that is resented by both speaker and hearer.

> RUINED SURPRISE. Taylor is planning a surprise party for their spouse, Tracy. Tracy accidentally overhears Taylor talking on the phone and comes to know about the surprise party.

In this case, much like the DISTRACTED STUDENT case, neither speaker nor hearer is intending for the transmission to take place, but unlike DISTRACTED STUDENT, neither speaker nor hearer *want* the knowledge transmission to take place. This kind of unintended and *unwanted* knowledge transmission demonstrates the futility of our epistemic situation; we can find ourselves on the receiving end of a knowledge transmission even when we explicitly intend not to.

Resented Knowledge Transmission

The third type of cases that can cause problems for Greco's shared intention principle involve the transmission of resented knowledge, that is, knowledge that the hearer does not intend to acquire and, if acquired, would be resented. Resented knowledge transmission cases are hard to categorise because they generally depend on individual preferences; some people would resent coming to know something while others would not mind.

These cases can be thought of in terms of *counterfactual conditional*s as the hearer is not in the epistemic position to know whether they would want to know whether *p* until they have already been transmitted knowledge that *p* or not *p*. Hearers in cases of resented knowledge transmissions might thus say "if I knew what you were going to tell me I would have asked you to keep silent", or, "if I knew how you would answer my question, I never would have asked in the first place". Here are four examples of resented knowledge transmissions that correspond to different emotions.[20]

[20] These cases seem prima facie to highlight that there can be ethical considerations involved when transmitting knowledge, but it is not clear that there is something epistemically problematic about them. To alleviate these

AFRAID. After your yearly check-up, the doctor tells you that you have an asymptomatic terminal illness and have only a few days left to live. You wanted to know whether you were healthy but after hearing the diagnosis you might have wanted to remain ignorant and live out the rest of your life without worry, but as soon as the doctor told you the diagnosis, the cat is out of the bag. This could be especially egregious when considering the (either real or imaginary) effect positive thinking can have in conjunction with treatment.

JEALOUS. You are at a restaurant with your partner celebrating your anniversary. Your partner then tells you how they used to go to this restaurant all the time with their previous partner and, in fact, they proposed to them at the very same table at which you are now sitting. Although this is not some horrific revelation, we can imagine some people would experience discomfort. One might say: "I did not want to hear this at the start of the date, it kind of killed the romantic atmosphere".

DISGUSTED. Someone tells you about Loa loa, a parasitic worm that travels under the skin of infected humans, across their eyes and into their lungs. This knowledge could prove useful to someone that is going to a rain forest in West Africa, but most people would rather go about their day without it.

ENVIOUS. A friend tells you that a mutual acquaintance, a former business partner of yours that betrayed you, is doing exceptionally well. After hearing this you find yourself feeling unhappy for the rest of the day.

Note that whether or not these cases exemplify UKT depends on the different personality traits and emotional states of the people involved; some people would not be bothered at all by any of these cases while others would experience them negatively. Still, we can accept that there

concerns, consider that even if the reason for not wanting to know something (and the reason one should first contemplate before transmitting knowledge that might have an adverse effect on the hearer) is not epistemically interesting, the process in which the knowledge gets transmitted is of great relevance. In sum, even if the reason for not wanting to know something is not epistemic in nature, the way in which one comes to know despite having an aversion to do so is epistemically intriguing.

generally exists some knowledge for any individual that they would not want to know for the sake of their emotional well-being.[21]

One line of argument that can be used to counter these examples is that people that experience negative feelings in these cases are emotionally immature, or at least less-than perfect epistemic agents, not using their cognitive faculties as well as they should have. Even if we concede that it is unreasonable to expect models to account for personal flaws of individuals, there is yet another kind of UKT, characterised by its achievement-depriving characteristics, that is still problematic for Greco's shared intention principle.

Achievement-Depriving Knowledge Transmission

Achievement-depriving knowledge transmission (ADKT) is a type of UKT that can disrupt an achievement-earning attempt[22] of a hearer. Spoiler transmission cases are archetypal examples of ADKT; they incorporate knowledge that, if transmitted, would be resented by the hearer because it would spoil their attempt at achieving something[23]. Consider the following cases of ADKT.

> CROSSWORD PUZZLE**.** You are solving a crossword puzzle and you are on the verge of completing it; you just need to find the last word. A friend of yours walks past you, sees that you are struggling, and proceeds to tell you the missing word before you can let them know you did not want any help.
>
> MATH PROBLEM. You are solving a difficult math problem to prepare for a test. You are halfway through the calculations and trying to figure out how to proceed. You are confident you will eventually figure out what to do next, thereby expanding your

[21] Note that there is a difference between not wanting to know that *p*, and not wanting *p* to be true.
[22] These can be both present and future achievement-earning attempts. If you were reading a book and someone spoiled the ending, then your relevant achievement-earning attempt would be ruined. If you were not reading the book, but you might want to read it one day, then your chances of ever earning that achievement would be severely reduced.
[23] Note that there is no need to define the exact nature of the achievement, it is enough that the hearer is averse to having their achievement-attempt sabotaged. This aversion is the motivating factor for being uncooperative in ADKT instances, which results in the hearer not sharing the speaker's intention and thus not a part of the relevant joint agency.

knowledge of how to solve these types of math problems. Your spouse, who happens to be a mathematician, sees that you are working on a math problem and tells you what you need to do to solve the problem, not realising you did not want any assistance.

MOVIE SPOILER**.** You are waiting in line to see the new mystery thriller movie when a stranger that just got out of the theatre walks up to you and tells you how the movie ends, including the plot twists that lead up to the shocking reveal.

These cases describe situations in which you want to earnestly try to find the answer yourself by competently using your own abilities, but someone spoils your achievement-earning attempt by telling you the answer. Moreover, if they intend to tell you the answer, there is nothing you can really do about it; you are not cooperating with the speaker (or, in some cases, the testimonial knowledge source) and yet you come to know the things you are being told.

In these cases, we find successful testimonial exchanges that transmit knowledge from a speaker to a hearer, but they do so without shared intention as it has been characterised. Specifically, we see that knowledge that *p* is transmitted from a speaker S to a hearer H, where S successfully tells H that *p*, S knows that *p*, S asserts that *p* with the intention of sharing knowledge that *p* with H, but H *does not* share S's intention even if H understands it, and they *do not* act jointly to bring about their shared intention (the intention of S to share knowledge that *p* with H). So, Greco's shared intention principle does not hold here and his reliance on joint agency is misguided.

One objection that immediately comes to mind is that speakers in cases like MOVIE SPOILER prove themselves to be untrustworthy by the testimonial act itself, as intentionally spoiling movies is surely not the mark of a trustworthy agent. If we assume that hearers try to cooperate with speakers, using their cognitive abilities to encourage knowledge transmission, then their competent agency would make them sensitive to this kind of untrustworthiness.

One way to address this objection is to argue that although the speaker in MOVIE SPOILER could be said to be *morally* untrustworthy there is nothing to indicate that they should be considered *epistemically* untrustworthy. It could even be argued that, because spoilers generally

need to be true for them to have the desired spoiler-effect, hearers in cases such as MOVIE SPOILER should be *more* inclined to believe what they are being told when they are being told something they do not want to know. Furthermore, in CROSSWORD PUZZLE and MATH PROBLEM, where the speakers do not act with malicious intent, there is nothing to indicate that they are untrustworthy (epistemically or morally).

Objections

What are the potential ways to defend against the cases presented here? There seem to be at least two distinct paths available. First is to argue that UKT cases are not actually knowledge transmissions, and Greco's shared intention principle is thus not compromised as it does not need to account for those cases. Second is to argue that UKT cases are not only cases of knowledge transmission, but that they fit within Greco's account.

If we were to argue that there is no knowledge transmission happening in cases of UKT, how would we go about it? We could say that UKT cases should not be considered cases of knowledge transmission because they involve a hearer that is not cooperating, which means that the hearer does not share intentions with the speaker and is therefore not competently partaking in joint agency. Sure, in some testimonial exchanges the hearer is relieved of some of the epistemic burden that usually comes with acquiring knowledge on your own, but that is not always the case.[24] If we think about a police investigator questioning a potentially uncooperative witness, we can see that even if the witness is telling the truth, they are not just passing knowledge on to the investigator in the same way they would when they tell their child that there is milk in the refrigerator.[25] Greco does not find this worrying, as he does not hold the view that all testimonial exchanges must transmit knowledge, even in cases where the speaker knows, and not even necessarily in cases where the hearer comes to know from testimony of a knowing speaker (Greco 2020, 5). The transmission thesis just claims that in some important type of testimonial cases, a speaker knows that *p*, reliably testifies that *p*, and a hearer comes to know that *p* because of the testimony of the speaker (Greco 2012, 21). Greco

[24] For further discussion on these different kinds of testimonial exchanges, see Greco (2012) and Lackey (2006).
[25] These examples are borrowed from Greco (2012, 33).

could even say that UKT cases describe situations in which the expectation of cooperation is inappropriate. He claims that condition (4) of his account of knowledge transmission, that S and H act jointly so as to bring about their shared intention, cannot be satisfied by speakers and hearers in situations where the expectation of cooperation is inappropriate[26], as the norms that govern cooperative testimonial exchanges are at odds with the norms that govern uncooperative testimonial exchanges (Greco 2020, 59). So, condition (4) only requires that the speaker and hearer act appropriately to bring about their shared intention and this is not possible in uncooperative situations (Greco 2020, 59).

However, this line of argument falls apart when we consider cases like CROSSWORD PUZZLE, in which the speaker and hearer can hardly be said to be uncooperative beyond the fact that the hearer did not like what they heard, but this is wildly different from the cases that Greco has in mind here (e.g., the police investigator). If cases like CROSSWORD PUZZLE were not to be considered knowledge transmission, instead just being examples of communicative exchanges, it would be difficult to find exactly what makes knowledge transmission special. As previously discussed, Greco's theory would become severely limited in scope.

Furthermore, even in uncooperative contexts, many UKT cases, such as RUINED SURPRISE and MOVIE SPOILER, look like typical cases of knowledge transmission. This result should not come as a surprise; when we examine the nature of transmission, regardless of the context in which it is applied, we see that transmissions generally do not require trust or cooperation to be successful; a virus does not require cooperation to cause a pandemic.

More generally, we see that if Greco were to bite the bullet and claim that UKT cases are not cases of knowledge transmission because they do not adhere to his account in some way, he would simply be begging the question by claiming that certain kinds of knowledge transmissions should not be considered as such because they do not conform to the conditions of his account that details the necessary conditions of knowledge transmissions.

[26] Greco states that the kind of trust that is appropriate for knowledge transmission can be inappropriate in contexts of knowledge generation and vice versa (Greco 2020, 59-60).

Now, a different route to defend against the problems presented in this chapter is to argue that UKT cases are compatible with Greco's account. One could say that the speaker and the hearer in UKT cases still share S's intention to tell H that *p*. Some actions can have both an individual action sense and a joint action sense, where the former refers to the actions of an individual and the latter the joint action of the participating actors. In acts of telling, like we have here, the joint action sense is the prevalent one, just like in speech acts of betting or promising. These unwanted knowledge cases are just exploiting this distinction by framing the knowledge transmission as being an individual action when it is not the case here. That is, even if the speaker intends to tell a hearer something, or more generally let them know about something, the hearer cannot share the speaker's intention as the hearer is not the one performing the action of telling or more generally, "letting know" (Greco 2020, 60).

This objection is one Greco indirectly considers in his book. He says that knowledge transmission *does* involve the kind of shared intention that joint action implies, because the speech act of telling, much like the action of betting or promising, is ambiguous in nature; these actions can refer to the individual actions of a single actor, or they can refer to their part of a joint action, in which they share an intention with someone else to do something together (Greco 2020, 60).[27] For example, when people make a bet there are two distinct actions in play, one is the individual action sense of trying to make a bet[28], and then the joint action sense of *actually* making a bet *with someone*, that is, the bettor has not really made a bet with someone until the bet has been accepted, which would not make sense if the individual action sense of betting was the only one (Greco 2020, 60). Furthermore, it seems like the joint action meaning of betting is the more commonplace one, which indicates that the individual actions in these cases are parasitic on the joint action meaning; if someone tries to make a bet and the other participant

[27] He gives an example of Brady throwing a ball into a practice net and how it is different from Brady throwing a ball to his teammate, Gronkowski. If the teammate does not catch the ball, then Brady merely tried to pass the ball, but was unsuccessful in his attempt (Greco 2020, 60). If Brady is to complete the pass, Gronkowski must catch it. We can respond by saying that this is not accurately reflecting what is really happening in cases of telling as a joint action. A more fitting example would be a case wherein Brady throws the ball as forcefully as he can towards Gronkowski's head with the aim of passing the ball to him (for him to pass the ball successfully it is enough that Gronkowski ends up with the ball in his hands as a direct result of Brady throwing it to him), Gronkowski grabs the ball instinctively so as to not get hurt. Brady has completed the pass, but Gronkowski did not share Brady's intention, and he would not be considered as a participant in a joint activity.

[28] An individual action sense of making a bet is, for example, when someone says: "I bet I can fix this.".

refuses the bet, the would-be bettor has not really made a bet (Greco 2020, 60). Greco argues that it is not even necessary to claim that the joint action meaning is the prevalent one, all that is needed is that there "exists the joint action meaning of "telling," on the analogy with a joint action meaning of "betting" and "promising"." (Greco 2020, 60).

However, there seems to be a stark difference between betting and telling; when a speaker intentionally spoils a movie for a hearer by telling them that Dumbledore dies it seems that the hearer "accepts" the bet automatically, as they come to know what they have been told and are aware of it when watching the movie for the first time, and it seems implausible to concede that they can refuse the knowledge transmission like they would refuse a bet (Greco 2020, 60). What this means in sum is that unwanted knowledge can be successfully transmitted without the joint action meaning of telling, and even if we concede that this individual action meaning of telling is parasitic on the more common joint action meaning, it still allows for the unwanted knowledge transmission to go through.

Another way to argue that UKT is compatible with Greco's view is to say that there is actually some sort of cooperation happening. One way to formulate this response is to say something along these lines: Knowledge transmission, regardless of whether the knowledge is wanted or not, involves a sort of fundamental cooperation, where both speaker and hearer are competent agents that are competently using their cognitive abilities to have a successful testimonial exchange that transmits knowledge. If this is true, then UKT cases adhere nicely to Greco's account.

We can respond to this objection by pointing out that this sort of fundamental cooperation would not be the primary factor in the hearer successfully acquiring knowledge. Even if the speaker and hearer would be indirectly credited with success by individually applying their cognitive abilities to competently cooperate, the success of the transmission cannot be credited to their *joint* competent agency as the necessary conditions for joint agency are not met. In other

words, as the hearer does not share the speaker's intention to share knowledge that *p*, they cannot act jointly to bring it about.[29]

We can modify condition (2) of Greco's account of knowledge transmission (that the speaker asserts that *p* with the intention of sharing knowledge that *p* with the hearer) in a way that allows Greco's shared intention principle (that state that the hearer understands and shares the speaker's intention and they act jointly so as to bring about their shared intention) to hold. A modified condition (2)* states that the speaker asserts that *p* with the intention to share knowledge with the hearer. Note that the speaker in (2)* does not intend to share knowledge *that p* with the hearer, just knowledge in general. This way, conditions (3) and (4) hold, as the hearer understands and shares the speaker's intention to assert *some* knowledge and they jointly act to bring about their shared intention of the speaker sharing knowledge with the hearer.

If we were to accept this modified version of Greco's account of knowledge transmission, then it would be reduced in a way that all that could be said of testimonial knowledge exchanges and joint agency is that the relevant intention is just to share knowledge in the most general sense, with no regards to the content of testimony.[30] Furthermore, this arrangement fails when we come to know something we did not want to know, as there is no standard mechanism in place for us to unlearn what we have come to know. Even worse, we cannot reliably filter out unwanted knowledge in advance without knowing the contents of said knowledge; the more you inquire about the content of what is about to be said the closer you are to knowing it.

[29] Note that joint causation does not imply joint agency. So even if two individuals both have an intention to walk to the same restaurant, they are not partaking in joint agency unless they are intending to walk to the restaurant *together*.

[30]. Note that this would result in Greco's transmission account being broadened in a way that would make it difficult to grasp what exactly is special about knowledge transmission. That is, if shared intention is modified to be inclusive enough to accommodate the idea that all communicative exchanges are considered joint activity, then, for one thing, at least some testimonial Gettier cases will be considered creditable achievements. I am grateful to an anonymous referee for prompting further clarification on this point.

If joint agency were to be diluted in this manner to account for the cases presented in this paper, then joint agency would overgeneralise to cases of knowledge transmissions that are obviously not consensual.[31]

Conclusion

The principal aim here has been to show that it is possible to transmit knowledge without shared intentions and show that Greco's account of knowledge transmission, which depends on these shared intentions through joint agency, is in trouble. Those who advocate for the sort of anti-reductionist knowledge transmission view that relies on shared intentions, like the view Greco presents, find themselves in a dilemma when they are put up against UKT cases. Either they concede that joint agency is not a necessary condition for knowledge transmissions, or they insist that it is present in all cases of knowledge transmissions, including UKT cases.

If they go for the first horn of the dilemma and say that UKT are not cases of knowledge transmission, then many paradigmatic examples of knowledge transmissions would not count as such, which in turn makes Greco's view only applicable to a seemingly artificially constructed subset of some such cases. This means that there will be many cases of knowledge transmission that fall prey to Lackey's dilemma, as there is no competent joint agency that can be credited for the true belief of the hearer.

If they go for the second horn and argue that UKT cases are cases of knowledge transmission, then they would be forced to accept a diminished form of joint agency that would simply be an ever-present by-product of testimonial exchanges. The triviality of this form of joint agency would make it difficult to discern what exactly makes knowledge transmission special and this account would overgeneralise to cases of knowledge transmissions that are obviously not cooperative and do not follow Greco's shared intention principle. This weaker joint agency cannot explain how a success can be credited to the hearer's competent agency even if that

[31] Note that this weakened form of joint agency also looks incompatible with a datum we find in the collective intentionality literature, namely, that praise and blame attributions can be appropriated to groups by shared agency. When shared agency has been stretched this thin it becomes unclear how group-level praise and blame would be appropriated as opposed to just individual-level praise and blame, as it is generally viewed as a function of cooperative interaction between the members of the group.

agency is not the most salient part of the success, as the success cannot be directly attributable to this trivial version of joint agency.

The ramification of the dilemma introduced here is that competent joint agency is either not present in many typical cases of testimonial knowledge exchanges, which makes those cases susceptible to Lackey's dilemma, or it is present in a diminished form that cannot produce the sort of shared credit that is needed to defend against Lackey's dilemma.[32]

[32] This paper was written as part of the Leverhulme-funded 'A Virtue Epistemology of Trust' (#RPG-2019-302) project, which is hosted by the University of Glasgow's COGITO Epistemology Research Centre, and I'm grateful to the Leverhulme Trust for supporting this research. Many thanks to Adam Carter, Chris Kelp, Mona Simion, Cian Brennan, Ross Patrizio, two anonymous referees, and the audience at the Scottish Epistemology Early Career Researchers WiP session, for helpful comments on earlier versions of this paper.

Chapter 2: Achieving Credit

Abstract. Knowing something seems better than merely truly believing that it is the case. But it turns out to be a challenging task to explain what makes knowledge distinctively valuable. Virtue epistemologists have presented an explanation using an individual-focused credit view of knowledge, which states that knowledge is a kind of achievement one gets by using one's cognitive abilities and that achievements have final value. This offers a promising way to escape our predicament. However, the view faces a well-known dilemma: any theory that aims to explain how we acquire testimonial knowledge makes it either too hard or too easy. I initially propose a social credit view, one that incorporates elements from speaker's accounts of knowledge transmissions. I then show how such views are challenged by a separate set of counterexamples in which hearers acquire knowledge from speakers that do not know. Finally, I present my *basic credit view* that emphasises that, in cases of testimonial knowledge, there must be a creditable epistemic achievement somewhere in the testimonial chain, a cognitive success that can be construed as having cognitive contact with reality. Objections that rely on unreliable epistemic agents, some of which target transmission views more generally, will then be systematically categorised and responded to.

Introduction

We feel accomplished when we come to know something because we satisfy our desire to discover the truth. But when we arrive at the truth about something by accident, this feeling of success dissipates. We seem to think that justified beliefs that are accidentally true are not as desirable as true beliefs that are justified in virtue of the employment of reliable cognitive abilities. This suggests that we care about something over and above truth—we care about how we arrive at it. Regardless, when we consider how our beliefs affect our actions in practice, it does not seem to matter how our beliefs were formed; the only thing that matters is that they are true. This question was raised in Plato's *Meno,* in which Socrates points out that true opinion is no less useful in action than knowledge (Plato, 98c), as having a *true belief* about the way to Larissa is just as instrumentally useful as *knowing* the way to Larissa. Even so, there is a kind

of value we pre-theoretically attach to knowledge that we do not attach to mere true beliefs.[33] The problem is pinpointing exactly where that difference lies.

This problem of explaining why knowledge is more valuable than mere true belief has also been called the *primary value problem* by those that argue that there exists (at least) a secondary value problem. The *secondary value problem* is that of explaining why knowledge is more valuable than anything less than knowledge (including justified true beliefs) (Pritchard 2007, 86-87).[34]

The way to argue against the primary and secondary value problems is to defend the notion that knowledge is more valuable than anything that falls short of knowledge, such as belief, true belief, or even justified true belief (JTB). But merely showing that knowledge is *more valuable* than whatever falls short of knowledge still suggests that knowledge and lesser-than knowledge are both situated along the same continuum. Having the difference between knowledge and anything lesser than knowledge be a matter of degree gives rise to other complications, at least for philosophers who think epistemology should primarily be concerned with knowledge, because then it would be unclear why we should be particularly interested in the exact point on the epistemic value scale that marks knowledge as opposed to any other point.

Discovering the allegedly unique value which only knowledge may deliver has been called the *tertiary value problem* (Pritchard and Turri 2014).[35] This last value problem, i.e., successfully showing why knowledge is qualitatively better than any lesser epistemic standings, can be stated in the form of the question; "Why is knowledge a more valuable *kind of thing* than that which falls short of knowledge?".

---

[33] For further reading see for example chapter six of Greco's *Achieving Knowledge* (2010, 91-101).

[34] Duncan Pritchard and John Turri (2014) provide an explanation of what this means. Suppose that knowledge is a JTB with an additional component that deals with Gettier-type cases and that justification adds value to a mere true belief. In that case, knowledge entails justification. It can then be argued that it is possible to answer the primary value problem because there is a property of knowledge that true beliefs lack that makes knowledge more valuable than true beliefs. This does not mean that we have a response to the secondary value problem. From these assumptions, we can see that JTB is a proper subset of knowledge, so the greater value of knowledge over mere true belief does not mean that knowledge is more valuable than any proper subset of its parts, including JTB.

[35] Viz. beyond being just a greater amount of value because it is further along the point of the scale than that which falls short of knowledge.

One way to address these value problems is to view knowledge as a *creditable achievement* one earns by exercising cognitive abilities. This type of a credit view can thus distinguish between accidentally true beliefs and true beliefs that arise from exercising one's reliable cognitive abilities. Having a credit view of knowledge would thus wield enormous explanatory power should it prove to be correct (Lackey 2009, 28). However, this virtue-theoretic view of knowledge has been criticised using counterexamples that involve cases of testimony, and as we will see, the attempts to respond to these criticisms have not been entirely successful.

Our starting point is that achievement is importantly tied to knowledge, and one way of cementing this importance is using a species of credit view that will be referred to here as the *individual* credit view. I will then introduce counterexamples that show that the individual credit view of knowledge cannot adequately account for cases involving testimony because it neglects the importance of the testifier and places undue responsibility on the person receiving the testimony.

In an attempt to save the general idea of credit views, I then propose a *social* species of a credit view that relaxes one of the requirements of the individual credit view, i.e., that it is always the hearer that deserves the credit in cases of testimonial knowledge. This social credit view is modelled after speaker's accounts of transmission theories and accounts for the testimony-based criticisms by only requiring there to be knowledge somewhere in the testimonial chain. Then, we will present counterexamples that involve the speaker transmitting knowledge they do not possess to show that even this social species of the credit view cannot account for all cases of testimony.

These counterexamples will then be used to motivate a minimalist species of a credit view, a *basic* credit view, which suggests that there are other achievements beside knowledge in testimonial cases. In short, we propose that cognitive success in the form of having cognitive contact with reality follows an achievement structure. The basic credit view is thus only committed to (1) that knowledge requires there to be credit somewhere in the testimonial chain, and (2) that the kind of creditable epistemic achievement that grants the credit does not need to be knowledge and can instead be a lesser epistemic achievement that is nevertheless a genuine epistemic achievement, namely, cognitive contact with reality through ability.

We then demonstrate that the basic credit view can address counterexamples aimed at speaker's accounts of knowledge. Specifically, it shows that someone in the relevant testimonial chain exhibits cognitive success by having cognitive contact with reality, and the hearer's knowledge is not without credit. Finally, we will examine further objections to transmission views in general that exploit the potential unreliability of epistemic agents and respond to them.

First Attempt – The Individual Credit View

The credit view of knowledge can be considered a type genus that is just committed to the idea that knowledge always and everywhere involves credit, i.e., that knowledge always arises from creditable belief formation. The standard species of this genus is an individual credit view, where the belief formation that is creditable when one knows must belong to that same individual. Some of John Greco's (2004, 2009, 2010) earlier work introduces and develops such a credit view that specifies that the individual thinker who attained achievement through ability deserves the credit. This version of a credit view can be seen as a defining example of an individual credit view. According to this earlier view of Greco's, knowledge is a kind of achievement an agent earns *in virtue of* their successful use of reliable and intellectually creditable cognitive abilities (Greco 2009, 224). When an agent successfully acquires knowledge by exercising reliable cognitive abilities, they deserve credit for *getting things right*. Furthermore, Greco argues that achievements are valuable for their own sake, which means that they possess *final value*. In short, achievements are finally valuable, knowledge is a kind of achievement, and knowledge is therefore finally valuable (Greco 2009, 225).

This entails that an agent's true belief is more valuable when it is a result of using reliable cognitive abilities than when the agent comes to believe something true by accident[36] because they deserve credit for arriving at their true belief by exercising their reliable cognitive abilities (Greco and Turri 2011). The individual credit view of knowledge plays a crucial role in virtue epistemological solutions to the value problems of knowledge because it can sufficiently

[36] Accidentally true beliefs are beliefs where the agent's cognitive abilities are not an important part of how they came to believe something (that happens to be true).

account for both the agent that possesses a true belief and the epistemic abilities[37] by which that true belief was formed (Greco and Turri 2011, Pritchard and Turri 2014).

Recently, Greco (2020), in his book *The Transmission of Knowledge*, has argued for an anti-reductionist theory of knowledge transmission that sees knowledge transmission and knowledge generation as two distinct phenomena, where one cannot be reduced to the other and each requires a distinct approach. So, in Greco's view, there are two ways of coming to know.

Firstly, individuals can generate knowledge from their competent agency (Greco 2020, 98). This competent agency is manifested by the individual when they competently use their abilities, such as their perception, to come to know something for themselves. As an example, imagine that you are standing in a field where a sheep walks up to you. Using your reliable faculties, you come to know that there is in fact a sheep standing before you based on your visual perception.

Secondly, individuals can receive transmitted knowledge by being a part of a competent joint agency (where both hearer and speaker are acting competently), where the true belief acquisition is attributable to the competent joint agency itself rather than the competent participation of the participating individuals (Greco 2020, 98-99). Greco's transmission view understands testimony as a function to distribute quality information already possessed by the epistemic community, rather than to bring quality information into the community for the first time (which would be more akin to the sort of testimony that facilitates knowledge generation in the hearer) (Greco 2020, 44).

According to Greco‘s recent view, knowledge generation is when one comes to know of their own accord, while knowledge transmission is when one comes to know from someone else without the usual epistemic burden that knowledge generation generally entails (Greco 2020, 1). Epistemic burden here can be thought of as the work required to generate knowledge by one's own efforts using one's own cognitive and sensory abilities. Knowledge transmission,

[37] Epistemic and/or cognitive abilities can also be referred to as virtuous sources in this context, see Pritchard and Turri (2014).

then, allows for a division of epistemic labour[38] in a way that knowledge generation does not (Greco 2020, 3-4). His argument, then, is that knowledge generation is to be understood in terms of success due to the competent agency of the knower. Knowledge transmission is to be understood in terms of success due to the competent joint agency of speaker and hearer acting together (Greco 2020, 19). Without complicating matters further, here is Greco's knowledge transmission view.

> **Greco's Account of Knowledge Transmission:** Knowledge that *p* is transmitted from a speaker S to a hearer H just in case S successfully tells H that *p*. And that happens just in case:
>
> (1) S knows that *p*.
> (2) S asserts that *p* with the intention of sharing knowledge that *p* with H.
> (3) H understands and shares S's intention.
> (4) S and H act jointly so as to bring about their shared intention (i.e., so as to consummate the speech act in condition 2) (Greco 2020, 57).

Greco's motivation here is to avoid Lackey's dilemma while still retaining a credit-based view of knowledge, thus being able to solve the value problems of knowledge without being susceptible to counterexamples involving testimony. However, it would be desirable to have a credit-based view that does not rely on joint agency, as it is unclear whether shared intention, a key component of competent joint agency and thus transmitted knowledge, is always present in cases of testimony (Ólafsson 2023). More importantly, there are counterexamples aimed at speaker's knowledge accounts of testimony, which involve testimonial knowledge being transmitted without the speaker having that knowledge. These cases are prima facie problematic for Greco's new transmission view as it relies on the speaker knowing that *p* in order to transmit knowledge and it is unclear whether not knowing that *p* restricts one from having the intention to share knowledge that *p*.[39]

Furthermore, it is worthwhile exploring whether we can introduce a species of a credit view that accounts for the various counterexamples without having to distinguish between two kinds

[38] This refers to Goldberg's (2011) conception of the division of epistemic labour.
[39] For example, imagine that I have *access* to baked cookies, but I do not wish to eat them; I can still share them with others who will. What is being highlighted here is that it is plausible that one can have access to information without believing it in a manner that makes it possible to still share that information with others.

of testimonial knowledge in which one kind can be reduced to knowledge of another sort while the other cannot (Greco 2020, 43). Even if we do not fully succeed (i.e., creating a unified account of transmission and generation), we might at least end up with a version of the credit view that can be integrated into Greco's idea on the two functions of testimony, i.e., a credit view that can account for credit in cases of knowledge transmission and generation without relying on shared intentions.

Our first attempt at solving the testimony-based problems of credit views will be to apply Greco's original credit view that we take to exemplify an individual credit view of knowledge. The view can be stated as follows.

> **Greco's Individual Credit View of Knowledge**: S knows that *p* iff S deserves credit for believing the truth regarding *p* because S's reliable cognitive abilities are an important necessary part of the total set of causal factors that cause S to believe the truth regarding *p* (Greco 2004, 127-128).

Greco's original credit view states that people know something if they have a true belief and an important part of how they arrived at their belief is their exercise of their own reliable cognitive abilities. The primary value problem is solved because knowledge as an achievement is *finally valuable* in a way that true belief is not. To understand what that means, consider how much someone would pay for a pen, and if that amount would change if they learned that Albert Einstein used that pen to write his paper on special relativity. Most people would look at the pen differently, and yet it is difficult to explain how the pen has changed; An extrinsic property grants the pen increased value, but it is still the case that the pen is properly valued for its own sake and thus valued non-instrumentally.[40]

The pen Einstein owned has value independently of the instrumental value it possesses in virtue of being a writing instrument (Rabinowicz and Rønnow-Rasmussen 2000, 35). So, when we have something that is finally valuable then we find value in the thing itself regardless of its practical utility, because it is not a means to an end (where the end is some other greater value),

[40] For further reading on intrinsic value, see e.g., *Recent Work on Intrinsic Value*, edited by Rønnow-Rasmussen and Zimmerman (2005).

it is itself an end. In much the same way, we find that a true belief with the extrinsic property of having been formed through reliable cognitive abilities (thus being a cognitive aspect of the more general notion of achievement) is more valuable than a true belief without any such properties (Pritchard and Turri 2014). The secondary value problem is solved because knowledge is more valuable than any of its parts, even when all the parts are added together, so a true belief from ability is not as valuable as a belief that is true because of ability.

Finally, we see that knowledge is distinctively valuable from anything that falls short of knowledge because it is a cognitive achievement and, as Greco states, achievements in general are finally valuable (Greco 2009, 225). Knowledge being distinctively valuable from anything that falls short of knowledge brings us a compelling answer to the tertiary value problem. The proposed solution of the individual credit view of knowledge is not without its criticism, and compelling arguments have been made against it.

A Problem for the Individual Credit View

Jennifer Lackey (2009) argues that the individual credit view of knowledge is undermined by cases involving testimony because they show that it is possible to know something without deserving credit for knowing it. If true, then the individual credit view of knowledge is compromised. One of Lackey's most famous examples, which she uses to support her argument, is the case of the Chicago visitor.

> CHICAGO VISITOR. "Having just arrived at the train station in Chicago, Morris wishes to obtain directions to the Sears Tower. He looks around, approaches the first adult passerby that he sees, and asks how to get to his desired destination. The passerby, who happens to be a lifelong resident of Chicago and knows the city extraordinarily well, provides Morris with impeccable directions to the Sears Tower by telling him that it is located two blocks east of the train station. Morris unhesitatingly forms the corresponding true belief" (Lackey 2009, 29).

According to Lackey, Greco's individual credit view is too strict, as it rules out knowledge in CHICAGO VISITOR. Even though Morris' abilities are *a part* of the set of causal factors that

give rise to his true belief, they are not a *salient enough part* of them.[41] Rather, it is the epistemic labour of the passerby that explains why Morris comes to know where the Sears Tower is. For Greco, the explanatory salience of a belief being true is relative to the interests and purposes operative in their subject's practical environment (Greco 2008, 428), and those interests and purposes ought to govern our evaluation of the knowledge claim (Greco 2008, 434). However, this does not threaten to make a theory of knowledge impossible by preventing our knowledge language from consistently identifying the same phenomena across different contexts (Greco 2008, 428). The reason is that the standards for knowledge will not differ significantly across various practical environments because of the functions that knowledge and knowledge language have in our practical and social activities (Greco 2008, 429).[42]

Greco's initial response to the CHICAGO VISITOR is that the case is underdescribed because what matters for evaluating whether a hearer's belief amounts to an achievement in testimonial cases is not the reliability of the speaker (or the hearer's knowledge of the speaker's reliability), but the hearer's reliability as it relates to receiving and evaluating testimony (Carter 2024, 12).[43] According to Greco, CHICAGO VISITOR can be understood in one of two ways, depending on which details are added.

The first way is to expand on the case so that Morris is exercising reliable abilities when he asks a passerby for directions (for example, add that he would not ask someone that was visibly drunk or dressed like a tourist). In that case, Morris is a reliable receiver of testimony, and as such we can attribute knowledge to him as well as credit for his reliable abilities being an important part of the explanation for his true belief.

The second way, which is arguably closer to what Lackey envisions the case to be, is to clarify that Morris does not exercise his abilities beyond some minimal threshold to engage in a testimonial exchange (e.g., by asking the first person he sees without discrimination), in which case he is not a reliable receiver of testimony. In that case his abilities would not be salient in

[41] And that this is the case is regardless of whether one is a reductionist or anti-reductionist about testimony.
[42] For further explanation of why this is the case, see Greco's (2008) paper *What's Wrong with Contextualism?*
[43] A more recent view can be found in Greco's (2020) *The Transmission of Knowledge.*

explaining the success of his belief, so he would not deserve credit, but neither would he come to know (Carter 2024, 12).

So either Morris gets knowledge and credit, or he gets neither, depending on how the case is constructed, neither of which is inconsistent with our expectations for how knowledge and credit should be attributed in CHICAGO VISITOR, and in neither case does Morris get one and not the other. Morris' reliability as a hearer does not cease to be salient because the passerby turns out to be reliable. If, however, Morris is not a reliable receiver of testimony, then the testimony will result in neither achievement nor knowledge (Carter 2024, 12).

It is not obvious that we do not come to know when we are not reliable receivers of testimony, if we suppose that testimonial knowledge is roughly as widespread as we pretheoretically attribute it to be, then we have to make sense of how we come to know in paradigmatic cases where we ask complete strangers for testimony with minimal exercising of abilities. However, even if we grant that there is neither achievement nor knowledge in cases where Morris is not a reliable receiver of testimony, we must take a closer look at the case in which Morris is a reliable receiver of testimony. The problem is that although Morris' abilities as a receiver of testimony would be reliable, it is not clear whether they would be a more salient part of the causal explanation for his belief than the abilities of the passerby. Explanatory salience of a causal contributor is context dependent, as it partly relies "on the interests and purposes operative in the context of explanation" (Greco 2008, 436). Our interests and purposes as knowledge exchangers are a part of the mechanisms that govern causal-explanatory salience, and according to those interests and purposes we value reliable information sources (Carter 2024, 13). With that in mind, it is not entirely clear that the interests and purposes of Morris would favour his abilities over the abilities of the passerby with regards to their salience in explaining Morris' belief (Carter 2024, 14).

If we try to argue that Morris still deserves *some* credit for exercising his abilities as well as he could, even if the abilities of the passerby ended up being a more salient part of the explanation for Morris' belief, in order to keep achievement and credit together. Then the standards of credit

would be lowered which in turn would make agents in Gettier-type[44] scenarios eligible to receive credit for their Gettiered justified true beliefs. To see this, consider the following Gettier-type case:

> DOCTOR. Bob went to his doctor, Alice, because he was not feeling well. After talking to Bob and performing some tests, Alice concludes that Bob has condition X and prescribes the appropriate treatment. Bob forms the belief that he has condition X after listening to his doctor's diagnosis. Unfortunately, Alice got two pages in one of her textbooks mixed-up when she was studying some years ago, which made her think that the symptoms for condition Y were the symptoms of condition X. Fortunately, Bob does indeed have condition X, but his symptoms are highly irregular and closely resemble the symptoms of Y.

In DOCTOR it looks like an important part of the explanation for his belief is either that his doctor got confused or that his symptoms are irregular. Greco notes that the level of explanatory salience can shift relative to different contexts, and this can include shifts in the salience of intellectual ability (Greco 2008, 421).[45] However, in this case it looks like the doctor's abilities are not salient enough to count as the cause of Bob's true belief in lieu of luck. If we grant Morris credit in CHICAGO VISITOR when it is not clear that Morris' abilities are a more salient part of the explanation for his belief over the abilities of the passerby, then we also allow Bob to receive credit for his belief in DOCTOR. Both examples involve a hearer that forms his belief by trusting someone else's testimony.[46]

If Lackey's claim is correct, the credit view of knowledge faces a dilemma; either the credit view of knowledge is stringent enough to systematically withhold credit in cases where an agent has Gettiered true beliefs (like Bob), in which case it is too stringent to give Morris credit in the

[44] These sorts of cases involve luck that interferes with one's knowledge acquisition, resulting in accidentally true beliefs that are still justified (Gettier 1963).
[45] Namely, it is the interests and purposes operative in the subject's practical environment that should govern our evaluation of the knowledge claim.
[46] And, arguably, Bob is in a better position to trust his doctor than Morris is in trusting a passerby.

Chicago visitor, or it is weak enough to grant Morris credit for his belief, in which case it also grants agents in Gettier-type cases credit for their Gettiered true beliefs.

Lackey concludes her article with a question that succinctly summarises the problematic nature of testimonial cases: "[H]ow does a testifiee, whose belief is true almost entirely because of the competence of the testifier, deserve credit for the truth of the belief that she acquires via testimony?" (Lackey 2009, 41).

Although Lackey's creditworthiness dilemma presents a serious challenge for the individual credit view of knowledge, and perhaps credit-based theories in general, we can attempt to save the wider family of credit views by proposing a social species of a credit view that can account for the type of social interactions prevalent in testimonial cases. The way to do so is to relax the requirement of the individual credit view that the credit must be attributable to the individual thinker that forms the belief, thus allowing us to attribute the credit to the speaker in cases of knowledge acquired from testimony.

Second Attempt – The Social Credit View

Testimony has been defined as the intentional transfer of a belief from one person to another.[47] The epistemological literature on testimony focuses on how knowledge, or justified belief, is acquired based on what we are told (Lackey 2011, 71). A lot of the things we claim to know are based on the testimony of others. We trust other people when they testify to us about things such as what the stars in our solar system are made of, how sound travels in waves, or the exact speed of light (Lackey 2011, 71).

There is something epistemologically distinctive about relying on the epistemic authority of others in cases of testimonial knowledge, which expands the possible ways we can defend testimonially acquired beliefs against criticisms (Goldberg 2006). The individual credit view is largely focused on the hearer, but testimonial cases also involve a speaker. One attempt to

[47] See e.g., Pritchard (2004, 326) and Faulkner (2006, 156).

respond to Lackey's Chicago Visitor case is to construct a credit view that emphasises the interpersonal nature of testimony by focusing more on the speaker's role in testimonial cases.[48]

The idea here is to explore whether a credit view can explain how hearers in testimonial cases come to know something without deserving credit for their knowledge. The individual credit view might not be able to account for some testimonial knowledge cases, but that does not entail that there is no credit present in those cases. If we move from an individual credit view to a social one, that emphasises the speaker's contribution, we can account for cases like CHICAGO VISITOR because we are then able to grant the speaker credit for having cognitive success through ability and testifying their findings to a hearer.

There are cases where a speaker in a testimonial case is transmitting knowledge that was originally transmitted to the speaker through testimony as well. This phenomenon can be analysed with the concept of a *testimonial chain*. A testimonial chain occurs when testimonial knowledge is transmitted between two or more epistemic agents, where a speaker testifies something to a hearer, who can then testify about it to someone else, thus becoming a speaker themselves. Testimonial chains can be short, for example when a speaker testifies to a hearer, but they can grow longer when a hearer then takes on the role of a speaker, transmitting knowledge to another hearer, and so on. The social species of a credit view presented in this section is a kind of a *speaker's knowledge account of transmission,*[49] which relies on the notion of testimonial chains. Goldberg (2006) suggests that testimonial knowledge grants the hearer the right to *pass the epistemic buck* and defer challenges back to the previous speaker once their own justification has been exhausted, and that this should be recognised as an essential feature of testimonial knowledge (Goldberg 2006, 134).[50] The general idea is that when a hearer, in cases of testimony, relies epistemically on knowledge that *p* from a speaker, the hearer acquires

[48] This is not a novel move, for example, Hinchman (2005), Moran (2005), and McMyler (2011) have explained the distinctiveness of testimonial knowledge by highlighting the interpersonal nature of testimony (Baker and Clark 2018, 179).
[49] Peter Graham writes that speaker's knowledge type accounts are so popular they must form a piece of our folk epistemology (Graham 2016, 174).
[50] Furthermore, this right of epistemic buck-passing right does not depend on any distinctively testimonial principle of epistemic justification, so even reductionists should be comfortable accepting it (Baker and Clark 2018, 179).

knowledge because the speaker transmits their knowledge that *p* to the hearer, and if the speaker does not have knowledge, they cannot transmit it (Graham 2016, 174).[51]

We can look at testimonial chains and see how a testimonial chain begins with an epistemic agent S_1 that used their own reliable cognitive abilities to arrive at the truth about something, and in turn testified her findings to an epistemic agent S_2. If agent S_2 then testified what they heard from S_1 to S_3, then S_3 would have arrived at their belief because the testimony of S_2, but their testimonially acquired knowledge would still rely on the reliable cognitive abilities the epistemic source S_1 exercised to arrive at their true belief.

A typical example of a testimonial chain can be seen when an agent uses their reliable cognitive faculties to arrive at a true belief (granting them knowledge according to the individual credit view) and then transmitting their findings to another agent who acquires the relevant knowledge through testimony. Consider the following example.

> HABANERO. A journalist working for The New York Times hears about an urban farmer that is selling habanero peppers from his backyard. Intrigued, the journalist schedules a meeting with the farmer to inquire about his operation. The farmer tells the journalist that he is selling habanero peppers on his farm every day from sunrise to sunset. The journalist, believing the farmer, writes an article that states that the farmer is selling habanero peppers from his backyard every day from sunrise to sunset. A few days later, a woman named Elsa is reading the article during her lunch break and forms the true belief that the farmer is currently selling habanero peppers at his farm.

In this scenario, a case could be made that Elsa is a reliable receiver of testimony because she used her reliable cognitive abilities to some extent—she decided to read *The New York Times* instead of the *National Enquirer*—and she would not have believed the article if she had any relevant defeaters of its contents; if it stated that the farmer was selling sentient habanero peppers then Elsa would surely dismiss the article. However, Elsa's cognitive faculties are not the most salient part of why she arrived at a true belief. The most salient part of why Elsa

[51] For discussions on epistemic dependence more generally see e.g., Pritchard (2015), McMyler (2011), Lackey (2008), and Graham (2000).

formed a true belief is that the farmer exercised his reliable cognitive abilities and then shared his knowledge with the journalist.

We can observe how the testimonial chain is formed. The farmer is the source of the knowledge that he will be selling peppers every day from sunrise to sunset. He transmits this knowledge to the journalist by telling her as much. The journalist then, being a reliable hearer, forms a true belief based on the farmer's testimony. She then writes an article that conveys her testimonial knowledge to everyone who reads it. Finally, Elsa reads the article and forms a true belief based on the reliable testimony of the journalist. Elsa's knowledge is a result of a testimonial chain that can be traced back to the farmer and his reliable cognitive abilities.

The idea of a testimonial chain as described here can be found in Faulkner's (2011) book *Knowledge on Trust,* in which he claims that the necessity component of Lackey's notion of the transmission of epistemic properties (TEP), a fundamental part of the belief view of testimony (Lackey 2008, 39), should be adjusted to account for such testimonial chains. According to TEP, a testimonial exchange involves a speaker's belief, along with the epistemic properties it possesses, being transmitted to a hearer (Lackey 2006, 434). Lackey presents necessary and sufficient dimensions to the transmission of epistemic properties thesis, which read as follows:[52]

> **TEP-N**: For every speaker, A, and hearer, B, B knows (believes with justification/warrant) that *p* on the basis of A's testimony that *p* only if A knows (believes with justification/warrant) that *p* (Lackey 2008, 39).

> **TEP-S**: For every speaker, A, and hearer, B, if (1) A knows (believes with justification/warrant) that *p*, (2) B comes to believe that *p* on the basis of the content of A's testimony that *p*, and (3) B has no undefeated defeaters for believing that *p*, then B knows (believes with justification/warrant) that *p* (Lackey 2008, 39-40).

[52] Graham (2016) points out that some version of the necessity requirement has been stated by e.g., Ross (1986, 62), Burge (1993, 486), Welbourne (1983, 302), and Audi (1997, 410), and variations of the sufficiency requirement can be found in e.g., Coady (1992, 223), Fricker (1987, 57), and Burge (1993, 477).

In what follows, I will be interested in the kinds of testimonial chains that originate in agents whose relevant belief was formed in accordance with requirements for the wider family of credit views of knowledge. For practical purposes, here is a fleshed-out stipulative definition of the kind of testimonial chains we have in mind:

> **Testimonial Knowledge Chain:** $S_1,\ldots,S_n$ form a testimonial chain with respect to a proposition *p*, $TC_p(S_1,\ldots,S_n)$, where S_{i+1} receives testimony that *p* from S_i for all $i \in N$ such that $1 \leq i \leq n-1$, and S_1 came to know that *p* without testimony in accordance with the individual credit view of knowledge.

These kinds of testimonial knowledge chains show the relation between the knowledge acquired by the agent that used reliable cognitive abilities to do so and those that subsequently receive that same knowledge through testimony. Note that S_n will be the final receiver of testimony in any knowledge-anchored testimonial chain. It will be useful going forward to name S_1, the first member of a testimonial knowledge chain:

> **Epistemic Source:** If S's belief that *p* is based on testimony, then its epistemic source is the first member S_1 of a testimonial chain $TC_p(S_1,\ldots,S_n)$. If S's belief that *p* is not based on testimony, then the epistemic source of S's belief that *p* is S.

Let us revisit HABANERO. Although Elsa knows when the farmer is selling habanero peppers after reading the article, she does not deserve credit for her knowledge. It is the epistemic source, the farmer, who deserves credit for using his reliable cognitive abilities to know when he is selling peppers, and then successfully transmitting his knowledge to the journalist who in turn transmitted it to Elsa. In other words, if the farmer would have lied to the journalist, either on purpose or accidentally, and said that the peppers can only be bought from dusk to dawn, Elsa would have formed a false belief.

Note that if the journalist would have observed people buying habanero peppers on the farm consistently from sunrise to sunset every day for months, she might come to believe that the farmer was selling habanero peppers every day from sunrise to sunset regardless of what the farmer would say to her. In that case, Elsa's belief could be credited to the journalist, who

becomes an epistemic source herself by using her own reliable cognitive abilities to arrive at the truth about something, eventually transmitting her knowledge to Elsa through testimony.

Elsa is incapable of discerning whether the relevant item of testimony is conveying truth or falsehood in this example, but that is generally not the case when her own reliable cognitive abilities are the source of her belief. However, when the epistemic source, other than Elsa herself, is completely justified and creditable in their knowledge which is expertly transmitted, then Elsa can know that thing just as well as if she would be the epistemic source herself. To see how, consider that the reliable cognitive abilities of the epistemic source, which are the most salient part of how Elsa arrived at her knowledge, would also be considered reliable cognitive abilities according to the standard credit view of knowledge.

Although the relationship between epistemic agents in a knowledge-anchored testimonial chain can vary, we see that for S to transmit knowledge that *p*, S must know that *p*. Using the terminology presented here, an epistemic source is required for knowledge transmission to occur.

Broadly stated, the species of credit view being proposed here is that an agent can acquire knowledge about something through testimony without deserving credit for it iff the epistemic source of the knowledge obtained it by using their reliable cognitive faculties in accordance with the individual credit view of knowledge. In that case, the epistemic source deserves credit for the agent's knowledge because the reliable cognitive abilities of the epistemic source are the most salient part of the total set of causal factors that cause the agent to believe the truth about something. This social type of credit view can thus be stated as follows:

> **A Social Credit View of Knowledge:** S knows that *p* iff the epistemic source of S's belief that *p* deserves credit for believing the truth regarding *p* because the reliable cognitive abilities of the epistemic source are a salient enough part of the total set of causal factors that cause the epistemic source to believe the truth regarding *p*.

In non-testimonial cases, the epistemic source of knowledge that *p* is the agent S. This social credit view of knowledge can account for those cases in the same way as the individual credit

view of knowledge can. To see this, let us denote S_E to represent S in cases where S is the non-testimonial epistemic source of their knowledge acquisition.[53] When substituted into the original formulation of the social credit view it can be expressed as follows to capture non-testimonial knowledge cases:

> **Non-Testimonial Application of the Social Credit View:** S_E knows that *p* iff S_E deserves credit for believing the truth regarding *p* because S_E's reliable cognitive abilities are a salient enough part of the total set of causal factors that cause S_E to believe the truth regarding *p*.

As we can see, this is just a recreation of Greco's individual credit view! Barring any problems, we can now show how the social credit view of knowledge allows us to reject the first horn of Lackey's creditworthiness dilemma which stated that the standard of credit needs be low enough for Morris to deserve credit, in which case the credit view becomes susceptible to Gettier-type cases. To see how this arm of the dilemma could be rejected, consider the following.

In the Chicago visitor case, we can assume that the passerby used his reliable cognitive abilities to acquire knowledge about the location of the Sears Tower, as he "happens to be a lifelong resident of Chicago and knows the city extraordinarily well" (Lackey 2009, 29). The passerby is therefore the epistemic source of the knowledge he transmits to Morris by testimony. The important part of how Morris came to know the location of the tower is the passerby's usage of his reliable cognitive faculties. Morris' knowledge of the tower's location can then be credited to the passerby.

Now, what happens if the passerby in the Chicago visitor case has Gettiered beliefs about the location of the Sears Tower? As in the Chicago visitor case, we assume that the passerby is the epistemic source of his belief regarding the location of the Sears Tower. If Morris' epistemic source, in this case the passerby, is to deserve credit for his true belief then his reliable cognitive abilities must play an important part in the explanation for how he arrived at his true belief

[53] This is an abstraction of a testimonial chain with one member, which is effectively a non-testimonial scenario. For the purposes of this chapter I am excluding edge cases involving self-testimony.

about the location of the Sears Tower. As it stands, the passerby's belief is accidentally true, and not true because of his reliable cognitive abilities.

If this social credit view holds, we can reject the first horn of Lackey's dilemma as the passerby cannot be said to know where the Sears Tower is according to the social credit view, and it appears to be impossible for them to transmit knowledge they do not possess to Morris. We can keep the standard of credit high enough to withstand Gettier-type cases while still granting credit where credit is due. But what if it would be possible for a speaker to transmit knowledge they do not have?

Problems for the Social Credit View

If it would be possible to transmit knowledge one does not have, then the social credit view would be in trouble, as it relies on the epistemic source (the initial speaker), to have knowledge (TEP-N), thus credit, for the hearer's belief. As it turns out, there are counterexamples that show how a speaker is able to transmit knowledge they do not have. These examples involve hearers who appear to be in an epistemically superior position to the speakers with regard to the facts about *p* (Greco 2020, 35). How could this come about? For one, the speaker might have a belief that acts as a misleading defeater with regards to *p* (Greco 2020, 35). In other instances, the speaker might be Gettiered, which precludes them from coming to know that *p* (Greco 2020, 35). In both types of scenarios, the hearer, being in a superior epistemic position relative to the speaker, is able to attain knowledge that *p* after hearing the speaker's testimony (Greco 2020, 35).

A paradigmatic case of this sort, which is supposed to show that a hearer can attain knowledge that *p* from a speaker that does not know that *p* through testimony, is Lackey's (2006, 2008) case of the creationist teacher:

> CREATIONIST TEACHER. A devoutly Christian fourth-grade teacher believes that creationism is true, and that the evolutionary theory is false. However, the teacher knows that there is an overwhelming amount of scientific evidence against both beliefs. They admit that they are not basing their beliefs on evidence but on the personal faith they

> have in their god. The teacher does not want to impose their religious convictions on her fourth-grade students, opting instead to present material that is best supported by the available evidence, which includes the truth of evolutionary theory. Consequently, while presenting a biology lesson she asserts that "modern-day Homo sapiens evolved from Homo erectus". The teacher neither believes nor knows this proposition but their students form the corresponding true belief on the basis of her reliable testimony (2006, 434-435, 2008, 48).

This case seems to show that one can transmit knowledge one does not possess. However, there have been many objections to this case. Among them is Faulkner's response that although CREATIONIST TEACHER is problematic for TEP-N it is only because TEP-N is not a good formulation of the transmission principle (Faulkner 2011, 73). In his view, CREATIONIST TEACHER shows that speakers can pass on what others know, and that all that matters is that someone in the testimonial chain possesses knowledge that can be passed on, i.e., having testimonially based knowledge means to have the epistemic standing of someone else to explain one's knowledge possession (Faulkner 2011, 73).[54] He thus reformulates TEP-N as follows:

> **TEP-N*:** Where A believes that *p* through uptake of S's testimony to *p*, A testimonially knows that *p* only if S knows that *p* or S's testimony to *p* is the end of a testimonial chain and some speaker prior to S in this testimonial chain knew that *p* (Faulkner 2011, 61).

For simplicity's sake, Faulkner proposes an abbreviated version:

> **Transmission principle for testimonial knowledge (TK):** Where A believes that *p* through uptake of testimony to *p*, A testimonially knows that *p* only if a prior speaker knew that *p* (Faulkner 2011, 62).

[54] For a detailed explanation of why this is important, see chapter 3 of Faulkner's (2011) *Knowledge on Trust* in which he argues that reductivists will have to overcome Moran's (2005) problem of intentionality if they are to reject transmission principles (Faulkner 2011, 65).

This principle is compatible with the social credit view and can account for CREATIONIST TEACHER. In Faulkner's view, the teacher is a non-knowledgeable conduit, in which they are simply passing along words of someone else.[55] In this case, the scientists and authors of the material on evolutionary theory. The teacher is thus the source of her student's testimonial belief, which amounts to knowledge, but she is not the source of knowledge. Instead, the teacher is connecting her student's belief to someone who knows that *p*. In CREATIONIST TEACHER we can imagine that the authors of the textbooks, the scientists, or even other teachers at the school could serve as the source of knowledge for the student's beliefs.[56] However, it is possible to intercept here and claim that such a non-knowledge conduit creates a gap in the testimonial chain, causing the knowledge transmission to fail regardless.

We can respond to this by pointing out that even if the teacher does not believe what she testifies, thus not knowing what she testifies, she can still have propositional justification for it. This propositional justification can then be transmitted to the students who, because the students believe the teacher's testimony, acquire doxastic justification for believing that modern-day Homo sapiens evolved from Homo erectus (Wright 2016, 304). As a result, the students come to know that modern-day Homo sapiens evolved from Homo erectus. A much stronger case can be found in Peter Graham's writings,[57] which can be stated as follows:

> FOSSIL. A creationist teacher does not believe in evolutionary theory, but still teaches it because she is a dedicated and responsible teacher. She develops a near expert understanding of evolutionary science and fossils. On a fieldtrip she discovers a fossil that proves that ancient humans once lived in the area, something which had been previously unknown to anyone. She does not believe that ancient humans lived there, yet she tells her students they did. The students believe her, and because of her commitment to teaching along with her expertise, she would not have said what she said

[55] Carter and Nickel (2014) label this as a kind of a hearsay, i.e., "the passing along of words of another without taking any independent view of the reliability of the message thereby conveyed" (2014, 146). Although the teacher decides to present the material that is best supported by evidence, she does so in a way where she is parroting rather than testifying (Carter and Nickel 2014, 146). When prompted, the teacher would not be able to sincerely assert that she finds the material she is teaching reliable because she believes it to be false.

[56] Note that Lackey here could say that students come to know because the teacher's words are reliable and not by inheriting the epistemic properties being transmitted from the source of the teacher's information (Carter and Nickel 2014, 148).

[57] See e.g., Graham (2000, 377, 2016, 176).

> if it were not true. Her assertion is thus a reliable indicator, and the students would not be easily mistaken when relying on her.

Fossil is a strong counterexample to any sort of a speaker's-knowledge account of testimony that relies on an initial knowledge source in a testimonial chain, such as the social credit view, because it shows that it seems possible to transmit knowledge without anyone earlier in the testimonial chain having that knowledge.

One way to defend against FOSSIL is to claim that FOSSIL exhibits knowledge, but not *testimonial* knowledge, because such knowledge would be dependent on the information channel, and in both these cases the hearers acquire knowledge from relying on their background knowledge of the connection the speakers have with reality (Graham 2016, 178).[58] However, as Graham notes, the children in FOSSIL respond to their teacher just as if the teacher believed what they testified. The point being that if children ever learn from dependence on their teachers, then the same goes for FOSSIL (Graham 2016, 178).

A case with a similar structure to FOSSIL is presented by Adam Carter and Philip Nickel (2014), which can be summarised as follows:

> GRANT SCHOLARS**.** Three particle physicists, A, B, and C believe in a religious tenet R. Each of them has been awarded grant by an organisation affiliated with their religion to do research with no strings attached. All three physicists decide to set up a sophisticated laboratory setting to demonstrate on scientific grounds that a certain particle φ, whose existence would provide compelling evidence against R, is not observed in ideal conditions. However, A does observe φ, who then proceeds to submit a paper detailing this discovery to Nature. The editor that receives the paper, who also believes in R, then sends the paper to the two most qualified experts on the subject matter, B and C, to be reviewed. B and C then proceed to repeat the experiment and verify the results, leading to the paper to be accepted. A creationist high-school physics teacher presents the results of A's paper, that φ exists, to their class. The students come

[58] See e.g., Audi (2006) and Fricker (2006).

> to know that φ exists and that R is false, but the creationist teacher, along with A, B, C, and the editor, do not believe what she just told the class (Carter and Nickel, On Testimony and Transmission 2014, 150-151).

In GRANT SCHOLARS and FOSSIL we cannot simply respond by saying that there was some prior speaker in the testimonial chain that knew what the students came to know through testimony. Not only is TEP-N disproven, but Faulkner's TK as well (Carter and Nickel 2014, 149).[59] We can still attempt a response that expands on our response regarding the knowledge gap in CREATIONIST TEACHER. Stephen Wright (2016) suggests that FOSSIL and GRANT SCHOLARS do not undermine the transmission of *justification*, and as justification transmission is more fundamental than knowledge transmission these cases are not successful against the most fundamental kind of transmission principles (Wright 2016, 307). In FOSSIL and GRANT SCHOLARS we see how, like in CREATIONIST TEACHER, propositional justification is transmitted down the testimonial chain until the students come to know because they believe the teacher's testimony, thus gaining doxastic justification for their beliefs. Unfortunately, as Carter and Littlejohn (2021) point out, this kind of a response is flawed in a few important aspects.

Firstly, this line of reasoning will not be easily accepted by those who find knowledge to be the fundamental epistemological concept, this includes not only proponents of knowledge-first epistemology but also those that accept the traditional view that knowledge is factorable into constituent components (including justification) but deny that this means that justification transmission is therefore more fundamental than knowledge transmission (Carter and Littlejohn 2021, 7.77).

Secondly, it is not clear that justification transmission involves *propositional* justification over *doxastic* justification. Wright (2016) has a response to this, namely, that at the core of transmission is the idea that justification transmission is a matter of the truthmakers for the proposition "the speaker has justification for what she says" becoming truthmakers for the proposition "the listener has justification for what the speaker says", and that this principle is

[59] Carter and Nickel (2014) call attention to Burge's 2013 postscript where he admits that TK and similar formulations are defeated by these kinds of counterexamples (Carter and Nickel 2014, 154).

framed in terms of propositional justification (Wright 2016, 298, Carter and Littlejohn 2021, 7.79). This response highlights the importance of propositional justification, but that alone is not a compelling reason to think that propositional justification is *more* fundamental than doxastic justification (Carter and Littlejohn 2021, 7.80).

Furthermore, even if we grant that propositional justification is at the heart of transmission, it is not clear that the non-believers in CREATIONIST TEACHER, FOSSIL, and GRANT SCHOLARS, are in fact propositionally justified in the propositions she testifies to her students. However, someone's reason for being propositionally justified for a belief must be available to them (Carter and Littlejohn 2021, 7.81). If the notion of availability is too lenient, then too many propositional justifiers are allowed through, but if there are any limitations placed on what counts as an available propositional justification, e.g., that a reason is only available provided one would be disposed to affirm it under some triggering conditions, then it is no longer clear why the teacher, in all three cases, is propositionally justified in what she says, seeing as her testimony contradicts her beliefs (Carter and Littlejohn 2021, 7.81).

Third Attempt – A Basic Credit View of Knowledge

We keep facing challenges when focusing on knowledge as the kind of epistemic achievement required for credit, and it looks like credit views in general have difficulty maintaining a connection between knowledge and credit without running into counterexamples. If we take a step back and abstract what it means for something to be epistemic, we find that, fundamentally, it has something to do with a connection to reality. Such a connection is an epistemic achievement and thus creditable, and fundamental to further epistemic achievements like knowledge and understanding. Furthermore, there is a precedent for regarding fundamental epistemic goals this abstracted way. Linda Zagzebski (1996), in *Virtues of the Mind*, suggests that knowledge can be defined as cognitive contact with reality arising from acts of intellectual virtue (Zagzebski 1996, xv).[60] Having knowledge thus entails being connected cognitively to the truth in a manner that is good, desirable, or important (Zagzebski 1996, 267). So, intellectual

[60] Note that we do not need to endorse a notion of intellectual virtues that aligns with Zagzebski's view to highlight cognitive contact with reality as a fundamental epistemic goal.

virtues are defined not by their ability to bring about true beliefs, but by their reliable contribution, in a higher-order way, to cognitive contact with reality. Within the domain of epistemic achievements we already have an implicit hierarchical structure where some achievements stand in different relations to knowledge on this hierarchy, and although cognitive contact with reality is further down the hierarchy it is not without value.

The plan here is to present a version of a credit view that can maintain that knowledge always entails credit without being challenged by counterexamples aimed at TEP-N. I argue that having cognitive contact with reality from ability is a cognitive success from ability, or a *basic epistemic achievement*, and that such achievements are creditable. Then, I argue that testimonial knowledge requires there to have been cognitive contact with reality through ability somewhere in the testimonial chain.

If correct, this would allow us to maintain the claim that there is no knowledge without credit in testimonial cases. In non-testimonial cases of knowledge, we retain an individual credit view. The more sophisticated and demanding forms of epistemic achievements, such as knowledge and understanding, require more than cognitive contact with reality, e.g., the formation of true belief through ability, which means that the credit that one deserves for such achievements is more than the base credit one gets for cognitive contact with reality. However, note that in all cases of knowledge, including GRANT SCHOLARS and FOSSIL, we have such *basic* credit, even if there is no *knowledge* credit in the testimonial chain up to the final receiver of testimony.

What GRANT SCHOLARS and FOSSIL have in common, and arguably what makes them such convincing counterexamples, is that even if the purported epistemic source of the testimonial belief does not themselves believe in the proposition being testified in the testimonial chain, they are still exhibiting some kind of a success from cognitive ability that looks like an achievement. As noted earlier, defenders of the speaker's knowledge account could try to respond to FOSSIL by pointing out the background knowledge a hearer has about the connection between the speaker's testimony and reality. Although this argument proved unconvincing, there is merit to the idea of highlighting the connection the speaker has with reality.

This realisation will be the focus of our third and final attempt at saving credit views from the problems of testimony. The individual credit view, on which S's knowledge that *p* requires that S has a cognitive achievement (success from ability) vis-à-vis *p*, was shown to run into a dilemma when introduced to testimonial cases like CHICAGO VISITOR. The social credit view, which tried to accommodate testimonial cases by simply granting the knowledge source credit, could not respond well to cases that involve knowledge transmission without the speaker having knowledge themselves. On top of that, it is not clear how the social credit view, which grants the speaker credit in cases where the hearer is not in a good position to exercise their abilities, would handle cases where the hearer *is* doing a lot of epistemic work,[61] because knowledge is qualitatively different from non-knowledge beliefs, and it would be difficult to split the credit between hearer and speaker without encountering the problems of allowing partial credit.

It looks like both the individual credit view and the social credit view are too demanding. The individual credit view requires the receiver of testimony to be creditable and the social credit view requires the epistemic source of a belief to be creditable, both of which fall prey to counterexamples that involve testifiers that do not know what they say. In light of all this, we might as well attempt to shift the focus away from belief. As we encountered in one of the responses to the schoolteacher cases, i.e., CREATIONIST TEACHER, FOSSIL, and GRANT SCHOLARS, it is possible to think about transmission in other terms than knowledge or belief, for example, propositional justification (Wright 2016). Imagine the following case:

> **CULTIST**. A man grows up in a cult that teaches him that everything is an illusion, and they are in fact in a matrix-like scenario. The man believes this,[62] and this belief serves as a defeater for nearly every other belief he could have. That is, he falsely believes that almost every belief he would have, regardless of whether beliefs can be voluntary, will turn out to be false. Consequently, he forms almost no beliefs about anything. With

[61] E.g., in Greco's police interrogator case where a police officer is interrogating a crime suspect that would be unlikely to divulge any information that might convict them (Greco 2020, 5). For further discussion on the continuum between second-hand (testimonial) beliefs and first-hand (non-testimonial) beliefs, see McMyler (2007). Graham (2016) notes that this point has received more attention in the literature on moral testimony, see for example Foley (2001), Gibbard (1990), Hills (2009), McGrath (2011), and Nickel (2001).

[62] Debating the effects of the cult's collective testimony being itself a part of the illusion is of little importance here, as what matters is that the man falsely believes that he is in a matrix-like scenario.

> nothing to be done, the man proceeds to live his life as he normally would while being a sort of a belief-zombie.[63]

In this case, the man is in a similar position to the teacher in CREATIONIST TEACHER and the scientists in GRANT SCHOLARS, where he has a belief that interferes with his ability to form other beliefs. Regardless, he can still acknowledge, just like the teacher and the scientists, that *if* he did not have his dogmatic belief, then he would be epistemically capable of discerning which beliefs he would form. The teacher in CREATIONIST TEACHER knows that, in the absence of a creator, evolutionary theory is best supported by evidence, which is why she decides to teach that to her students. The scientists in GRANT SCHOLARS know that, in the absence of their creationist belief R, they would believe the science, which is why they decide to submit their findings to *Nature*.

In both cases, the non-believers are playing an insincere game, in which they think "imagine my belief is false, what would then be the right belief to have?". By compartmentalising their belief in such a way they are fully capable of recognising good beliefs from bad ones, e.g., the teacher in FOSSIL would not be indifferent about what she would tell her students as a result of finding the fossil, and the scientists would not try to publish a paper filled with scientific errors.

In much the same way, if the man in the cult was a prominent physics researcher he would be capable of recognising good physics from bad physics, even if he does not believe that *our* physics represent the truth.[64] What this example shows, is that regardless of belief one can still exhibit cognitive achievement by successfully applying one's epistemic faculties, even if it is carried out under the guise of the "false" belief that one's cognitive achievements exist within an artificial frame.

As the counterexamples show, knowledge is not a prerequisite for knowledge transmission, but it looks like a cognitive success is required somewhere in the testimonial chain for knowledge

[63] This is borrowed from Robert Kirk's (1974) notion of a philosophical zombie, although Keith Campbell (1970) made a similar argument, later popularised by David Chalmers (1996).
[64] Although he is likely to have some beliefs such as 2=2, he is not in a position to know whether complicated physics theorems are correct, given how much of our current understanding depends on how *our* world operates.

to arise. I propose the following basic credit view, which preserves the idea that there cannot be knowledge without credit by offering a novel approach to understanding testimonial cases and retaining Greco's individual credit view in non-testimonial cases:

The Basic Credit View of Knowledge:

S knows that *p* either non-testimonially or testimonially.

If non-testimonially, S knows that *p* if, and only if, S formed a true belief with respect to p because S's belief that *p* is a result of cognitive success from ability (Greco 2010, 71).

Or,

If testimonially, S knows that *p* only if S formed a true belief with respect to p and there is an epistemic achievement in the form of a cognitive success (i.e., cognitive contact with reality) from ability somewhere in the testimonial chain.

The testimonial chain specified here can be defined as follows:

Testimonial Chain: $S_1,\ldots,S_n$ form a testimonial chain with respect to a proposition *p*, $TC_p(S_1,\ldots,S_n)$, where S_{i+1} receives testimony that *p* from S_i for all $i \in N$ such that $1 \leq i \leq n-1$.

This basic version of the credit view can withstand the specific criticism of both GRANT SCHOLARS and FOSSIL, namely, that one can come to know without there being credit. Note that the basic credit view is named that way because it is basic: it does not make the stronger claim that when someone in the testimonial chain deserves basic epistemic achievement that knowledge will inexplicably follow. However, it provides us with a new way to assess testimonial knowledge cases. This explains why I opted for Greco's individual credit view in non-testimonial cases. Having a disjunctive account allows us to rely on the more robust

individual credit view in non-testimonial cases while also maintaining that knowledge entails credit in all cases where there is testimonial knowledge.

In the typical creationist teacher case, we have argued that the original authors of the material being taught can be seen as the sources of the transmitted knowledge, even if the creationist teacher does not believe what she says, instead acting as a non-knowledgeable conduit. However, even if the original authors of the material were themselves non-believers, making the case akin to FOSSIL and GRANT SCHOLARS in the sense that there is no speaker in the testimonial chain that believes the relevant proposition, we still have an answer.

In FOSSIL and GRANT SCHOLARS, we see that even though the speakers do not believe what they say they are still performing good epistemic work. The scientists in GRANT SCHOLARS are expert physicists doing rigorous research and the teacher in FOSSIL is relying on their expertise to evaluate new evidence and making epistemically good inferences from it. Even so, without belief, they do not possess knowledge about what they say. For example, the scientists in GRANT SCHOLARS that found the particle do not exhibit the kind of achievement that knowledge entails, because they do not have an epistemic achievement insofar as they do not have a true belief from ability. Lacking true belief from ability, their cognitive success as it relates to knowledge does not manifest belief forming competence. However, they clearly do have cognitive contact with reality, and that cognitive contact is not a matter of luck. They make cognitive contact with reality because of their exercise of abilities. We defined this kind of an achievement as a *basic epistemic achievement* because the scientists' cognitive contact exhibits achievement structure, i.e., a success from ability. A basic credit view only requires there to be a basic epistemic achievement somewhere in the testimonial chain, of which such a cognitive success from ability falls under. Our view can thus reconcile how the hearers acquire knowledge that *p* in GRANT SCHOLARS and FOSSIL while the speakers do not know that *p*. The speaker's assertion that *p* is based on the reliable cognitive success from ability of someone in the testimonial chain that facilitated cognitive contact with reality, i.e., the cognitive connection to the truth that *p* (Graham 2016, 177).

One worry is that the basic credit view makes is too lenient, that is, it makes it possible for an agent to get credit for a failed attempt at coming to know. Our response to this worry is that

although there is a special value to knowing something, there is also some epistemic value in mere true belief because it is action guiding. If we consider that we are navigating the world as intelligent beings, there are at least two dimensions to our intelligence. On one hand we can score epistemic achievements by fitting mind to world (and when that goes well, we have knowledge or true beliefs), and then we have achievements in action when we fit world to mind (according to our desires). When considering the former kind of achievement, cognitive contact with reality remains a genuine epistemic achievement, even if it is only cognitive contact with reality through ability. We can concede that such an achievement is not as valuable as knowledge, but it is still the sort of achievement we should care about, because without those lesser kinds of epistemic achievement we can no longer reliably act on accurate information, seeing as how we would not recognise cognitive contact with reality, a fundamental part of further epistemic achievements.

Problems for the Basic Credit View of Knowledge

We are now in a position to address what could be the most plausible type of objections against transmission theories more generally (Greco 2020, 35). These objections involve a discrepancy between one's reliability as a believer on one hand, and the reliability of their actions in a testimonial exchange (viz. the actions that pertain to speaking and hearing) on the other (Greco 2020, 35). Here, reliability can be understood as follows. When an agent is reliable *as a speaker* (testifier), then they try to testify in a reliable manner. When an agent is reliable *as a hearer* (testifiee), then they generally comprehend testimony reliably. When an agent is a reliable *believer*, then they form beliefs in a manner that is consistent with good epistemic practices (whether that be through their own reliable abilities or reliable testimony).

A basic counterexample to transmission theories, which leverages the discrepancy between an agent's reliability as a believer and their reliability as it relates to testimonial acts, involves a speaker that tells a hearer the truth but would have said the same thing even if it were false. In such a case the speaker is a reliable *believer* but an unreliable *testifier*.[65] To illustrate, consider

[65] Goldberg notes that speaking of the reliability of testimony is a simplification that is sometimes appropriate but not always, because there are cases where the speaker is reliable about one aspect of their testimony while failing to be reliable in another (2001, 526).

a speaker that believes that *p* only if *p*, but they would always testify that *p* regardless of whether *p* is true.

The nature of these cases give rise to various permutations that can highlight different facets of testimonial exchanges. As a result, we find various versions of these cases in the literature.[66] To further expand on these permutations, consider that we can have a typical case of successful testimonial exchange that involves a reliable testifier that is also a reliable believer, who testifies to a reliable testifiee that is also a reliable believer. However, the speaker could be an unreliable testifier, or they could be a reliable testifier while being an unreliable believer, and the same goes for the hearer. Speakers and hearers can thus be evaluated along two dimensions here, reliability of testimony and reliability of belief forming. Note that a speaker that is twice unreliable (as a testifier and as a believer) can, in some cases, appear to be a reliable testifier. For example, if a speaker is an unreliable believer such that they always believe the opposite of what they hear, but then they always testify the opposite of what they believe. However, both dimensions (i.e., testifying and believing) are still unreliable when viewed independently, even though the unreliability of both dimensions can allow for scenarios in which one unreliable process cancels out the other, thus making the unreliable testifier that is also an unreliable believer *consistent* in their testimony.

It is clear that there is no unified response that can address all of these case variations, so they will be reviewed systematically. We will employ a three-character notation scheme where the first character identifies the participant (S for speaker, H for hearer), while the second and third characters denote their reliability as a testifier/testifiee and as a believer, respectively (R for reliable, U for unreliable). For example, SRU refers to a speaker who is a reliable testifier but an unreliable believer, while HRR denotes a hearer who is both a reliable testifiee and a reliable believer. A case involving such a speaker and hearer would thus be abbreviated as SRU-HRR. We can further define SRR and HRR to be *fully* reliable agents.

In total, there are sixteen case permutations involving these properties:

[66] See e.g., Graham (2000, 2006), and Lackey (2006, 436-437, 2008, 53-54).

- Speaker and hearer are both fully reliable**:** SRR-HRR.
- Speaker is fully reliable, but the hearer is unreliable: SRR-HRU, SRR-HUR, SRR-HUU.
- Speaker is unreliable, but the hearer is fully reliable: SRU-HRR, SUR-HRR, SUU-HRR.
- Speaker and hearer are both unreliable: SRU-HRU, SRU-HUR, SRU-HUU, SUR-HRU, SUR-HUR, SUR-HUU, SUU-HRU, SUU-HUR, SUU-RUU.

We can see that SRR-HRR are just standard cases of testimony that do not pose specific problems to knowledge transmission theories. Furthermore, any cases of knowledge that involve a hearer that is an unreliable believer (H*U) will not result in counterexamples to transmission theories, as the hearer's true beliefs will not be true because of any cognitive success but luck. We have already encountered SRU-HRR[67] structures in CREATIONIST TEACHER, FOSSIL, and GRANT SCHOLARS, and shown that although they are problematic for the social credit view, they do not pose a serious challenge for the basic credit view.

What we are left with then, are SUU-HRR, SRU-HUR, SUR-HUR, SRR-HUR, SUU-HUR, and SUR-HRR. These variations target transmission theories more generally, so even if the basic credit view can defend against SRU-HRR counterexamples, it is not obvious that it can defend against counterexamples that target transmission more generally, as the basic credit view relies on transmission just like the individual and social credit views.

We will divide the remaining cases into two categories, based on how we can respond to them. In the first category, which we can call *two-instance unreliability*, we have various paradigmatic counterexamples that involve two instances of unreliability. These include consistent liar (SUU-HRR), consistent testimony (SRU-HUR), and consistent miscomprehension (SUR-HUR). In the second category, we have *single-instance unreliability*, we have cases that involve

[67] SRU-HRR also includes Lackey's persistent believer, which shows how knowledge that *p* can be transmitted to a hearer when the speaker believes that *p* without knowing that *p* or having justification for their belief that *p*. The basic credit view can account for such cases in a manner that is akin to how it accounts for schoolteacher cases, namely, that the speaker in "persistent believer" has cognitive contact with reality because of their abilities, even if they do not have the relevant knowledge or justification. A similar case can be seen in Goldberg (2005) where a speaker with an unsafe belief testifies to a hearer but, because of an onlooker that would intervene if the speaker testified falsely, the belief the hearer forms after listening to the unsafe testimony is safe (2005, 302).

only one instance of unreliability, but the testimony is unsafe. These cases include Lackey's (2006) dishonest whale-watching business owner and Graham's (2016) hospital case (both of which fall under SUR-HRR), as well as SRR-HUR and SUU-HUR[68].

Two-Instance Unreliability

SUU-HRR

These cases involve a speaker that is both an unreliable testifier and an unreliable believer and yet appears to be capable of producing reliably true testimony. A famous portrayal of such a case is Lackey's consistent liar case, which can be summarised as follows:[69]

> CONSISTENT LIAR. Bertha suffered a head injury when she was a teenager which resulted in a brain lesion that made her prone to lying, particularly about her perceptual experiences that involve wild animals. A doctor tried to operate on the lesion but discovered that it was impossible to repair, as a last resort he created another lesion that (1) made Bertha's pattern of lying extremely consistent, and (2) would combine in a very precise way with a pattern of consistent perceptual unreliability. The doctor did not tell anyone about this change of plans. As a result, Bertha is a radically unreliable but highly consistent believer with respect to her perceptual experiences of wild animals. For example, whenever Bertha sees a deer, she believes that it is a horse. However, she is also radically and consistently insincere, so nearly every time she sees a deer and consequently forms the belief that she saw a horse, she insincerely testifies to others that she saw a deer. Because she is so consistent in her beliefs and intended lies, no one has any reason to doubt her reliability as a source of information and she is in fact considered to be one of the most trustworthy people in her community. While talking to her neighbour Henry she reports insincerely but accurately that she saw a deer on a nearby

[68] Note that SUU-HUR involves three instances of unreliability, yet it can pose similar challenges as SUR-HRR and SRR-HUR when structured like CONSISTENT LIAR and the hearer having single-instance unreliability.
[69] Graham's case of the wine taster, modified from Dretske (1982), is similar in that it portrays a false connected belief (Graham 2000, 370-371). Another case that involves speakers that do not believe, or have justification, for what they say, involves a twin Earth scenario where colours are inverted and the words that refer to colours are similarly inverted (Graham 2000, 379-380).

> hiking trail (while believing she saw a horse). Henry, having no relevant defeaters in addition to finding Bertha very trustworthy forms the true belief that there was a deer on a nearby hiking trail (Lackey 2008, 53-54).

If Henry has testimonial knowledge that there was a deer on the hiking trail, then even the basic credit view is in trouble because it seems that neither Bertha nor Henry is in cognitive contact with reality, and yet it seems Henry comes to know that there was a deer on the hiking trail. An initial line of response is to bite the bullet and deny Henry knowledge regardless. However, that would also mean that much of the testimonial knowledge we think we have is not knowledge, as the way we think we acquire testimonial knowledge is indiscernible from the way Henry does. To see why, consider that Lackey's dilemma applies here just as in the CHIGAGO VISITOR case, that is, Morris could just as well have been talking to a consistent liar. The only difference is that Bertha is generally considered to be trustworthy unlike the passerby in CHICAGO VISITOR, making CONSISTENT LIAR the stronger of the two cases.

A second line of response is to claim that Bertha is not actually testifying because of their confused state, and because there is no testimony we can claim that Henry ends up with non-testimonial knowledge. This is a weak argument. Whether it be confusion or maliciousness, Bertha is in a position to lie, which suggests that she is also in a position to testify. Even so, the idea of non-testimonial knowledge can be employed for a third, more promising, response.

There are cases where a speaker testifies to a hearer, but the hearer does not receive testimonial knowledge. I suggest this is one of those cases. Consider that if Henry knew about Bertha's condition he could then work out when she is telling the truth e.g., whenever Bertha talks about a deer Henry can reason that she thought she saw a horse but is lying about it by replacing the horse with a deer, and furthermore, if Henry read Bertha's personal diary and she wrote that she saw a horse standing in a field yesterday, he could confidently say that she saw a deer yesterday. Now, Henry might consider her less trustworthy in general, but he would still be able to reliably acquire knowledge from Bertha even though she is not in cognitive contact with reality. Benjamin McMyler (2011) says something similar when he points out that if Henry were to learn that Bertha's testimony is insincere but true then Henry might still use Bertha's testimony as a guide to the truth (even acquiring knowledge), but he might nevertheless stop

trusting Bertha (McMyler 2011, 85). He goes on to say that without trust in Bertha it is unclear whether this counts as a straightforward instance of testimonial knowledge acquisition, and if testimonial knowledge and belief involve trusting others, then Henry cannot acquire testimonial knowledge from Bertha (McMyler 2011, 85-86). Another way to frame this is to say that Henry's belief would not be *testimonially* based, but *instrumentally* based, as he would be treating Bertha's testimony as he would a clock to determine the time. When Henry relies on a clock to know what time it is, he comes to know the time because of his background information about the clock (e.g., he knows that it is calibrated with real time reliably enough), and he can rely on Bertha in the same way, as he possesses background information that Bertha is a trustworthy testifier.

Sanford Goldberg (2012) argues that both testimonial-based and instrumental-based belief cases involve epistemic reliance on an information source (Goldberg 2012, 217). In both cases there is "a state of affairs" involving the information source that the belief-forming agent takes as the source representation of what is the case. When the information source is a speaker in a testimonial case the output is their performance of a particular speech act in which a proposition is presented-as-true, but when the source is instrumental, the output is the state of an instrument being regarded by the belief-forming agent as a representation of what is the case (Goldberg 2012, 217-218). The agent's belief-formation in both cases is guided by the semantic content of the representation and the epistemic ground for belief, where the epistemic grounding refers to forming a belief on the basis of accepting the representation (Goldberg 2012, 218). Goldberg finds that when such representation's outputs and processes are not "appropriately subject to normative epistemic assessment", then we should not regard them as "relevantly extended" belief-formations, but instrumental (Goldberg 2012, 220). In contrast, when one relies on an epistemic agent which is themselves appropriately subject to normative epistemic assessment, then that would be testimonial.

In our current case, we see that Henry does not rely on Bertha's normative epistemic assessments, but on the brain lesions that make her testimony reliably true. The testimony itself does not play a role in explaining the knowledge, so Bertha's act of testifying would be epiphenomenal.

If we accept that Henry's knowledge is not based on Bertha's testimony, then the basic credit view can account for it. Namely, Henry's non-testimonial knowledge is creditable because he has cognitive contact with reality through ability, analogous to how he would be creditable for having cognitive contact with reality when using his ability to read the time from a clock.

SRU-HUR

In these cases, like consistent liar, we have two instances of unreliability where the unreliable factors coincidentally cancel each other out. The difference is that both speaker and hearer are unreliable in one of two ways, instead of the speaker being unreliable in both. Consider the following case involving a speaker that is a reliable testifier but an unreliable believer, testifying to a hearer that is an unreliable testifiee but a reliable believer:

> CONSISTENT TESTIMONY. Darl is unable to reliably form beliefs about animals, but because of a brain lesion he always tells the truth about what animal he saw without being aware of what he said (i.e., when he talks about an animal he saw, he hears himself say the name of an animal species but does not make any connections between the word he just said and any particular animal). In this case, he is talking to Cash, who coincidentally also has a brain lesion that makes him unable to reliably hear testimony correctly when people are telling him about animals. To Cash, the word horse and the word deer sound remarkably similar, so much so that he cannot distinguish between them. However, he still reliably forms true beliefs from such testimony because the brain lesion also makes him subconsciously and preternaturally sensitive to micro-expressions, contextual clues, and so on.

In this case, it does not seem difficult to deny Cash *testimonial* knowledge, as the most salient reason for why he forms true beliefs is his preternatural ability. To further illustrate, consider the following example that is relevant to CONSISTENT TESTIMONY as well as the CONSISTENT MISCOMPREHENSION case in the next section:

> BRAIN CHIP. Natalie gets a brain chip implant that turns any testimony she listens to, true or false, into a truth. Whenever someone speaks to her, the chip overwrites what

> was said, causing Natalie to hear a random truth instead. The truths the chip transmits to Natalie always correspond to beliefs that the relevant speaker has. This is a highly reliable process, and Natalie has never been given false information by the chip.

If Natalie were to be asked where her knowledge comes from, it would not be from the incoherent or insincere testifiers. Even if the lies and the mumbling *cause* the brain chip to trigger, those actions are not the basis for Natalie's beliefs. The difference is that Natalie's true beliefs are not exactly caused by her reliable abilities, but the point of BRAIN CHIP is to emphasise that these cases do not involve testimonial knowledge. The hearers in CONSISTENT LIAR and CONSISTENT TESTIMONY do not possess testimonial knowledge because their beliefs are only reliably true because of a head injury and heightened sensitivity, respectively. BRAIN CHIP further shows why we excluded SUU-HUR, as the hearer's state would the same with the only change being that the speaker would be a consistent liar (SUU), making it even clearer that if the hearer acquires knowledge at all, then it would be non-testimonial knowledge. The basic credit view can account for CONSISTENT TESTIMONY just like CONSISTENT LIAR, as Cash's non-testimonial knowledge is a creditable epistemic achievement because they exhibit cognitive contact with reality through their reliable (preternatural) ability.

SUR-HUR

These cases involve a speaker and a hearer who are both reliable believers but unreliable in their testimonial acts. Graham (2016) presents such a case involving identical twins that only talk to each other:

> CONSISTENT MISCOMPREHENSION. One of the twins (S-Twin) consistently lies about their perceptual judgments of wild animals because of a head injury. The other twin (H-Twin) is operated on in secret which results in an altered state which causes H-Twin to perceive S-twin's utterances about wild animals in a particular manner. Namely, when S-Twin sees a deer and believes it is a deer, they assert that they saw a horse, but H-Twin understands S-Twin's testimony such that S-Twin is asserting that they saw a deer. H-Twin thus comes to believe that there was a deer close by. When H-Twin forms

> this belief, they would not easily be mistaken, they reliably form a true belief by relying on their representation of what S-twin asserted (Graham 2016, 179-180).

In this case, H-Twin is not a reliable receiver of testimony (because they do not have the correct uptake of S-Twin's testimony), but because of the specific way S-Twin is unreliable in their testimony, H-Twin's belief forming process is reliable. We can respond to this case in a similar fashion as we responded to CONSISTENT LIAR and CONSISTENT TESTIMONY. Namely, that if H-Twin acquires knowledge, it is not testimonial knowledge, and the basic credit view can grant credit appropriately because H-Twin relied on their reliable abilities and background information to arrive at their true belief.

Single-Instance Unreliability

SUR-HRR[70]

These cases involve a speaker that is a reliable believer, and as such has cognitive contact with reality, but their testimony is unreliable. One such example can be found in Lackey (2006):

> WHALE. A dishonest whale-watching business owner, who has financial incentive to say that whales have been sighted in the area regardless of whether that is the truth, is asked by a potential customer whether there have been whale sightings and replies truthfully. However, the business owner would have said that there had been whale

[70] Note that SRR-HUR can also be used to construct cases such as WHALE and HOSPITAL. To illustrate, in SRR-HUR we have a speaker who is fully reliable and testifies to a hearer that is generally an unreliable testifiee (bad hearing, for example), but they luckily hear what the speaker testified and consequently form a true belief. The problem is that they could very easily have misheard the speaker and formed a false belief. If we alter the case so the speaker is a consistent liar, then we have created a case of SUU-HUR, where the two instances of speaker unreliability turn them into a reliable source of information, with the single-instance unreliability of the hearer being problematic because they are lucky that they heard the consistent liar's testimony correctly. However, because SRR-HUR and SUU-HUR do not produce examples that are interestingly different from the SUR-HRR cases such as WHALE and HOSPITAL, we will only respond to SUR-HRR, as the same response can be applied to SRR-HUR. SUU-HUR elicits a slightly different response. Although it can be structured similarly to WHALE and HOSPITAL, it will be less compelling, as CONSISTENT LIAR cases arguably do not involve testimonial knowledge.

> sightings regardless of whether there were any whales spotted in order to sell more tickets (Lackey 2006, 436-437).

A similar example can be found in Graham (2016), who modified the case from Nozick (1981):

> HOSPITAL. A father knows that his son is fine today, even though he suffers from medical issues. The father tells his mother (the son's grandmother) the truth, viz. that her grandson is fine. However, the father would have said that his son was fine even if he were not, as to not upset her. The grandmother would easily form false beliefs by relying on her son's testimony (Graham 2016, 175).

These cases are set up as counterexamples to the sufficiency condition TEP-S of speaker's accounts of knowledge, as the speaker (the business owner in WHALE and the father in HOSPITAL) know that *p*, but the hearer does not come to know that *p* on the basis of their testimony (Graham 2016, 175).

These cases are not counterexamples to the basic credit view. The conclusion of HOSPITAL is that the father knows, but the grandmother does not because of insensitivity. The testimonial branch of the basic credit view is not committed to saying otherwise, because it only claims that a speaker somewhere in the testimonial chain must have creditable cognitive success, in the form of cognitive contact with reality, *in cases of knowledge*. The basic credit view only claims that when a hearer acquires testimonial knowledge, there is credit somewhere in the testimonial chain. It does not argue for the position that if there is credit somewhere in the chain there is necessarily knowledge. HOSPITAL and WHALE are cases where the final hearer does not acquire knowledge, even if there is credit somewhere in the chain.

Conclusion

The basic credit view, if correct, provides a novel way of defending the pretheoretically plausible idea that credit always follows knowledge, saving credit views from having to concede the connection between the two.

We showed why the individual credit view is vulnerable to testimonial cases like CHICAGO VISITOR. We then introduced a distinctly social type of credit view that is inspired by speaker's accounts of knowledge, and demonstrated how FOSSIL and GRANT SCHOLARS oppose such a view. Namely, those counterexamples show how knowledge that *p* can be transmitted to a hearer without the speaker knowing that *p*.

Finally, we presented a basic credit view, which relies on the notion that there are basic epistemic achievements that might not be as valuable as knowledge or understanding but are nevertheless genuine achievements that are creditable as such. Having cognitive contact with reality from ability is one such achievement, as it is a cognitive success from ability that follows an achievement structure. On the basic credit view, testimonial knowledge requires there to have been cognitive contact with reality through ability somewhere in the testimonial chain. If correct, this allows us to maintain that there is no knowledge without credit. In cases of non-testimonial knowledge we refer to the individual credit view.

We acknowledge that the more demanding forms of epistemic achievements, such as knowledge and some cases of understanding, require more than cognitive contact with reality, namely a true belief from cognitive ability, and the credit for such achievements is greater than the basic credit one gets for cognitive contact with reality from ability. The basic credit view shows that there is no knowledge without credit, as the individual credit view, which the basic credit view maintains in non-testimonial cases, provides an explanation for why knowledge is incompatible with intervening epistemic luck in non-testimonial cases, namely, that an agent can only be ascribed knowledge if their true belief is because of cognitive success from ability. In cases of testimonial knowledge, someone in the testimonial chain exhibited some cognitive success from ability, which is a creditable basic epistemic achievement. If we accept this, it looks like GRANT SCHOLARS and FOSSIL do not pose a threat to a fundamental claim of credit views, viz. that knowledge always entails credit.

Chapter 3: Temporal Elements of Trust

Abstract. There are some underdeveloped features of trust that can provide a meaningful distinction between certain instances of trust that have often been seen as being interchangeable. These features are temporal in nature and have so far not been correctly accommodated for in some paradigmatic accounts of trust. Firstly, a distinction will be made between trust affirmations, which is the initial trusting relation that is formed, and ongoing trust, which is the succeeding trusting relation. Secondly, a distinction will be made between definite and indefinite trust which differ in the way they relate to trust resolutions, i.e., how time affects the entrusted action, and it will be shown that successful therapeutic trust cannot be a case of indefinite trust. I then make a novel distinction between disclosed and undisclosed monitoring. The concept of confirmation monitoring will be presented, which will be shown to be a necessary attribute of therapeutic trust and, more generally, not in tension with definite trust. Then, I evaluate precautionary measures of trust as they relate to the risks of betrayal and disutility. Finally, three accounts of trust and monitoring will be analysed using these temporal elements of trust to show that they cannot account for these temporal elements in some instances, while being beneficial in others. The accounts in question are Arnon Keren's doxastic preemptive reasons account, Emma Gordon's beneficial monitoring account, and Wanderer and Townsend's account of trust and rationality.

Introduction

Trust is fragile (Baier 1986, 260), but it is nevertheless a commanding experience. We can quickly assess whether a situation involves trust, whether we trust someone, and how strong that trust is. Even so, it has proven to be difficult to fully capture the nature of trust. Philosophers have attempted to confine trust within a framework of their choosing, whether that be doxastic[71], non-doxastic[72], performance-theoretic[73], or pluralistic[74].[75] As a result, we are left with a variety of trust accounts that all seem plausible to a degree while being somewhat incompatible with one another. Among the reasons for trust being difficult to grasp is that our

[71] See e.g., Adler (1994), Hieronymi (2008), McMyler (2011), and Keren (2014).
[72] See e.g., Baker (1987), Holton (1994), Jones (1996), Faulkner (2007), and McLeod (2002).
[73] See e.g., Carter (2020, 2022).
[74] See e.g., Simpson (2012), Scheman (2020), and McLeod (2020).
[75] Further and separate distinctions can be made, e.g., motive-based vs. non-motive-based.

intuitions vary between instances of trust. For example, it is intuitive to think that trusting someone is a reason to believe what they say, while also intuitively thinking that acting on evidence is at odds with trust. I believe some of these disagreements can be explained and reconciled by identifying various temporal elements of trust that have until now been underdeveloped within epistemology.[76]

These predominantly overlooked temporal features of trust differ between paradigmatic cases of trust, which becomes problematic when those cases are then evoked to argue for various trust accounts without accounting for their temporal differences. The resulting theories of trust are thus left with a narrow conception of trust that disregards temporal effects, which results in inconsistent outcomes when faced with other trust cases that seem identical when their temporal properties are not accounted for.

The elements of trust that will be elucidated and categorised in this chapter are temporal, i.e., the effects of time on trust instances. Although there are many interesting questions to consider, for example whether time passing can *ceteris paribus* weaken or strengthen trust, this chapter will only focus on a few distinct elements of trust. We begin the chapter by pointing out that there is a meaningful difference between the trust that gets established initially, and the trusting relation that follows. We will define these terminally distinct trust acts as *trust affirmations* and *ongoing trust* respectively.

Then, we will distinguish between *definite* trust and *indefinite* trust in three-place trust cases. Namely, that definite trust cases have a set time limit based on the action or inaction one is trusted with, while indefinite trust cases have no such time limit.[77] Although these two distinct types of trust have often been used interchangeably, we will show that they are affected differently by precautionary measures, such as monitoring.

---

[76] Some philosophers have used temporal elements as they relate to trust, such as Edward S. Hinchman (2021) when he writes: "*If we view an intention as an intrapersonal trust relation that unfolds through time between distinct selves, earlier and later, which self invites trust and which accepts the invitation?*" (Hinchman 2021, 84). Others, like Kenny (1963), Vendler (1957), and Jespersen (1924), have identified and analysed temporal elements within the philosophy of language (particularly as they relate to verbs).

[77] Although indefinite trust cases do not have a terminus analogous to definite trust cases, they can still have a different kind of a time limit, e.g. when a monogamous couple breaks up then their trust that they will be faithful to one another comes to an end.

After making these distinctions, we are in a position to argue that one kind of precautionary measure, namely a specific type of monitoring we will define as *confirmation monitoring*, is essential to definite trust cases, and an important part of trust building in therapeutic trust cases. This conclusion is antithetical to the common conception of trust within the literature, namely that any and all precautionary measures tarnish trust to some degree.

Following that conclusion, we demonstrate that not only have the temporal elements of trust been underappreciated, but that the discourse on risk in relation to trust has often been unclear, as it tends to conflate two distinct types of risk that are not always differentiated in the literature. We thus distinguish between two kinds of risk: the consequentialist risk of disutility, and the deontological risk of betrayal. This risk classification can contribute to our understanding of precautionary measures, their purpose, and the subtle differences between them. Some objections will be considered when relevant and are used to further strengthen the view presented here. Finally, three prominent trust accounts will be examined in light of these temporal elements of trust. Namely, Arnon Keren's (2014) account of preemptive reasons, Emma Gordon's (2022) account of beneficial monitoring, and Wanderer and Townsend's (2013) account of trust and rationality. We will argue that all three accounts are affected when we direct our attention to the time schemata presupposed by various instances of trusting.

Trust Affirmations and Ongoing Trust

In a typical three-place trust example A trusts B to Φ, A relies on B to Φ, and A would feel betrayed if B would not do as entrusted,[78] where trusting someone to Φ is successful iff they Φ as entrusted, and not merely iff they Φ (Carter 2024, 130). Some philosophers claim that A needs to believe that B is trustworthy, or at the very least that B will Φ.[79] Others claim that trust does not require belief, but is instead a matter of attitude or disposition.[80] Yet others claim that trust is a performance, where the focus is on how the attitude on trusting is normatively

[78] To do something as entrusted can mean e.g., to Φ with goodwill towards the trustor (Jones 1996, Baier 1986), to Φ with the interests of the trustor in mind (Hardin 2002), or to Φ because one believes they have a commitment to the trustor to Φ (Hawley 2014).
[79] See e.g., Adler (1994), Hieronymi (2008), McMyler (2011), and Keren (2014).
[80] See e.g., Baker (1987), Holton (1994), Jones (1996), Faulkner (2007), and McLeod (2002).

constrained.[81] Although we will assume a doxastic[82] account of trust here, in the sense that trusting someone to do something at least entails a corresponding belief, the main interest here will be the intention and action of trusting, how temporal elements can affect trust cases, and time-sensitive monitoring.

These typical three-place trust examples rarely focus on the Φ itself, i.e., the action that is involved in the trusting relation. What kind of actions can B be trusted with? When we examine the plethora of trust cases that have been somewhat representative of trust, we find that the entrusted actions are not uniform. Instead, we find that the entrusted actions can generally be split into two categories. In some instances, Φ can be something like "A trusts B to bring them something", in other cases "A trusts B to honour their arrangement for the unforeseeable future". Even though these different kinds of actions have no obvious bearing on the nature of three-place trust relations, they still need to be adequately accounted for if we are to end up with a complete picture of trust.

The first step to untangle these different kinds of actions is to make a distinction between *trust affirmation* and *ongoing trust*. Trust affirmation is the trust that the trustor initially finds themselves to have towards the trustee Φ'ing, and it arises from the trustor's belief or disposition that the trustee is trustworthy. B's trustworthiness is established by the trustor's previously held beliefs or attitude towards the trustee, based on the trustor's beliefs and experiences. As it is defined here, trust affirmations need only be performative insofar as to let the trustee know they are trusted in order to establish a three-place trusting relation between the trustor and trustee and thus avoid cases of mere reliance. It should be emphasised that the trust affirmation itself does not suffice for trust to obtain in the way someone saying "I promise" suffices as a promise. It is not a voluntary speech act, although such an act may entail to further establish a trusting relation, rather, trust affirmations are the state of one's trust at the exact moment one involuntarily finds that one trusts. When the trustor affirms their trust in the trustee, there is no room for monitoring or rational reflection. The trustor trusts the trustee in that

[81] See e.g., Carter (2020, 2022).
[82] Proponents of doxastic accounts are committed to two claims, firstly, that one does not trust a person at all if one does not believe them to be trustworthy and secondly, that the more you doubt someone's trustworthiness, the less you trust them (Hieronymi 2008).

moment in time without opportunities to undermine it.[83] Consider that one's trust can diminish with time, so although one fully trusted at the time of the trust affirmation, that trust does not always persist; I can trust you with my car and then get second-thoughts and consequently stop trusting you with the car, even though you are still driving it.

The idea of a trust affirmation intuitively seems like a standard example of two-place trust, where the trustor has a trusting attitude towards the trustee in a way that does not involve a specific action from the trustee that the trustor relies on. So, at the onset of trust, one could simply state that at the time that trust has been established between the trustor and the trustee, then that trust affirmation is a simple two-place trust that would then turn into a three-place trust if and when the trustor asks the trustee to Φ. Although intuitive, this notion of trust affirmation would be misrepresenting the scope of the idea. Trust affirmation as it will be defined here is relevant to both two-place and three-place examples of trust.[84] In both cases of trust, there is a distinction to be made between the moment that trust is established (either towards the trustee as being trustworthy in general, or trustworthy to Φ), and the trust that follows, which we will define as ongoing trust.

In cases where the trustor trusts the trustee to be faithful to them, the trust affirmation seems less significant than what happens afterwards. Namely, the ongoing trust that emerges from the affirmation. This kind of ongoing trust is much more susceptible to undermining by means of precautionary measures such as monitoring, rational reflection, and searching for counterevidence.

For clarity's sake, let us define trust affirmation and ongoing trust as they appear in three-place trust relations:

[83] One could try to argue that some instances of trust are just this kind of trust affirmations with no ongoing trust to follow, for example, some cases of speaker-trust. When B testifies that *p* to A with the intention of having A trust what they say, and A in fact trusts B that p, then the transaction is complete; B cannot reasonably presume that A will believe that *p* indefinitely. As long as A trusts B during the initial testimony, B is satisfied that A trusts them. However, when we consider a more general case of trust with similar time-sensitive properties, such as a climber that grabs the hand of another climber, these cases certainly look like examples of a very fast ongoing three-place trust with limited opportunities to take precautionary actions.

[84] One could argue that, if one-place trust has a defined starting point, then one-place trust affirmation makes sense.

> **Three-place Trust Affirmation:** At time t_0, A initially forms the belief that B is trustworthy and trusts that B will Φ.
>
> **Three-place Ongoing Trust:** At time t_1, after A affirmed their trust that B will Φ at t_0, A continues to trust B to Φ until t_2, at which point B has either performed or failed to perform Φ.

Ongoing trust is a standard three-place trusting relation that is either definite or indefinite and requires trust affirmation as a starting point. To better understand what ongoing trust entails, especially at t_2, we need to shift our attention towards the different forms of trust, namely, definite and indefinite trust. Definite cases of trust are aimed at a specific action at a specific time, while indefinite trust cases involve trust without a set terminal point.

Definite Trust and Indefinite Trust

Trust affirmation and ongoing trust are two temporally distinct stages of trust, but there is another underappreciated property of trust that can vastly differ between instances of trust. Recall that our three-place ongoing trust definition states that A continues to trust that B will Φ until t_2, at which point B has either performed or failed to perform Φ. Now, what exactly occurs at t_2 after such a three-place trust has been established between A and B? There is a significant difference between cases in which A trusts B to Φ, and the Φ is never epistemically *confirmed* by A to have been executed by B (and in trust cases that involve actions that do not have a terminus it cannot be confirmed), and cases in which A trusts B to Φ, and the trust is eventually confirmed after the action has been executed by B.

We can define trust confirmation as follows:

> **Trust Confirmation:** A trusts B to Φ at time t, B executes Φ at time ≥t, and A confirms their trust in B if A comes to know that Φ has been executed by B. If A will never be in a position to know whether B executed Φ, then A's trust cannot be confirmed.

A significant detail here is that sometimes trust cannot be epistemically confirmed, but B has still done as entrusted. Without confirmation there is no way to distinguish between successful unconfirmed cases of trust and failed unconfirmed cases of trust. This results in an undesirable situation where we cannot distinguish between instances in which A is unable to confirm their trust and B has done as entrusted, and instances in which A is unable to confirm their trust and B failed to do as entrusted. To address this, consider the following definition of trust resolution:

> **Trust Resolution**: A trusts B to Φ at time t, and the trust is resolved iff B executes Φ at time ≥t because of their commitment to A. If Φ does not have a terminus, and therefore cannot be executed at any time ≥t by B, then A's trust cannot be resolved.

Trust resolution is therefore not a matter of confirming whether the entrusted action was performed as a commitment to the trustor, but rather a matter of whether the trustee has performed the action they are entrusted with. One takeaway here is that trust failures will always be cases of unresolved trust.

From the perspective of the trustor we find confirmation to be an internal matter while resolution is an external one. We can now define definite and indefinite trust which will be the main subject of this section:

> **Definite Trust**: Three-place trust in which A trusts that B executes Φ, and the trust can be resolved.

> **Indefinite Trust**: Three-place trust in which A trusts that B executes Φ, and the trust cannot be resolved.

These definitions are useful to better grasp the temporal differences between different kinds of Φ. In definite trust, we have actions that have some kind of a terminus. For further clarification, consider the difference between trusting you to close the gate tonight (definite trust) on one hand, and trusting you to be the person who closes the gate each night on the other (indefinite trust). Taking care of things as entrusted in the former case involves performing an action that, once taken care of, resolves the trust because you have done as entrusted. In the latter case

however, there is no particular action at any particular time that can resolve the trust. In this respect, the former case is akin to running a race while the latter case amounts to something like "to continue running".

Only definite trust proceeds towards a terminus, which explains why it is relatively easy to confirm definite trust cases (did they perform the action of running the race?), while it seems implausible one can epistemically confirm indefinite trust cases (did they perform the action of continuing running?). Meanwhile, indefinite trust cases are cases in which the ongoing trust has no terminus and cannot be resolved or confirmed, such as when a couple trusts each other to be faithful. There is no exact point in time where the couple finds that the action that they trusted each other to perform has been completed even though they are considered to be trusting each other during any substretch of time.[85]

This discussion is not completely novel. We find discussions on temporal elements in the philosophy of language, where considerations that involve the concept of time is relevant to verb usage. Vendler (1957) distinguished between verbs without continuous tense, and verbs with continuous tense.[86] He divides the former category further into achievements and states, while the latter category contains a distinction between activity terms, that include e.g., "running" and "pushing a cart", and accomplishment terms, such as "running a mile" and

[85] One might argue that if the couple were to split up, without the trust being broken then that would constitute a kind of definite trust. However, it depends on the specifics of what the trusting action initially entailed. If the trusting relation was established as "we will never be unfaithful to one another", then, in case of a breakup, there is no way for either of them to break the trust or keep their word after the split. The trust is resolved only insofar as they did as entrusted until the conditions of their trust became trivial. Now, if the trust affirmation was "we will never be unfaithful while we are together", then this could potentially constitute ongoing definite trust when the relationship has a set time limitation that is known from the start. This raises a couple of questions. Firstly, whether instances of ongoing indefinite three-place trust instances can change over time, or a new trust instance takes the place of the old one. Secondly, how trust accounts can account for the human condition, i.e., mortality, in cases where a partner passes away. Just imagine someone on their death bed trusting a close friend to make certain funeral arrangements. This kind of trust affirmation cannot become ongoing, and the trustor has no way to monitor the trustee or follow through with their part of the trusting relation. Does this indicate that people that are close to dying are unable to trust others because they cannot take precautionary measures (even if they wanted to), they cannot be harmed by trust failures (as the realised risk cannot affect them posthumously), and they cannot be betrayed? One line of response would be to simply ascribe hope in lieu of trust to cases like these. Another response would be to point out that people often say things like "their legacy was betrayed", or that "their dying wish was not honoured", which at least indicates that there is some posthumous risk involved.

[86] A similar classification system can be found in Anthony Kenny's (1963) book *Action, Emotion, and Will*, which he developed independently of Vendler (Mourelatos 1978, 416). Kenny's framework did not separate achievements and accomplishments as distinct types like Vendler, which results in a trichotomy of activities, performances, and states (Mourelatos 1978, 416).

"drawing a circle" (Vendler 1957, 146). We see that the subcategories of continuous tense verbs closely align with the categories of indefinite trust and definite trust, as both category pairs share an intrinsic duration (Mourelatos 1978, 416).[87] Additionally, the differences we find between definite trust and indefinite trust mirror those between the subcategories of continuous tense verbs (activities and accomplishments). Activity terms, like indefinite trust, go through time in a homogenous way in which any part of the process is of the same nature as the whole, while accomplishment terms, like definite trust, go through time while "proceeding toward a terminus which is logically necessary to their being what they are" (Vendler 1957, 146). Continuing with the analogy between temporal verb-types and temporal elements of trust we find that, on one hand, both activities and indefinite trust call for periods of time that are not definite because they involve no culmination (Mourelatos 1978, 415). On the other hand, accomplishments and definite trust imply the notion of definite time periods (Vendler 1957, 149).

Alexander P. D. Mourelatos (1978) finds that the Kenny-Vendler typology is too narrowly conceived and that the trichotomy of activities, performances, and states is a part of a broader ontological trichotomy of processes, events, and states (Mourelatos 1978, 422). Under Mourelatos' scheme, indefinite trust would be categorised as a process, while definite trust would fall under the events category. He further classifies achievements as punctual occurrences and accomplishments as developments, with both categories belonging to the category of events (Mourelatos 1978, 423).

Finally Vendler's category of achievements captures the inception (or the climax) of an act, which can be dated, but cannot occur over a temporal stretch. Trust affirmations would thus be considered achievements in Vendler's time schemata, but more importantly, trust confirmations would be considered achievements (Mourelatos 1978, 416). This is significant when we consider Mourelatos' claim that there cannot be an accomplishment without a closely related end-point achievement, i.e., A cannot say that they "trust B to Φ" if they cannot eventually say

[87] Jespersen (1924) makes a distinction between verbs where the action is either confined to a single moment (such as "catch") or it implies a final aim (such as "make"), and verbs that denote an activity "which is not begun in order to be finished" (such as "love") (Jespersen 1924, 272-273). He defines the former class as conclusive and the latter as non-conclusive, which harmonises well with our concepts of definite and indefinite trust.

that they "trusted B to Φ" (Mourelatos 1978, 417). This suggests that there cannot be definite trust without a closely related trust confirmation.[88]

In elaborating on the differences between activity terms and accomplishment terms, Vendler (1957) perfectly sums up the difference between definite and indefinite trust as he writes: "Somehow this climax casts its shadow backward, giving a new color to all that went before", where the climax can be reinterpreted as the definite trust confirmation (in case of the trustor), or the definite trust resolution (in case of the trustee) (1957, 146).

The entrusted actions in definite trust cases are often isolated one-off actions that need not be performed again. But what about trust scenarios in which the task at hand is in some sense a "one-and-done" kind of task, but is performed repeatedly? One such task would be as follows:

> MORNING COFFEE. A trusts B to bring them coffee to bed every morning.

Is the kind of trust in MORNING COFFEE definite or indefinite? One might respond by saying that the task both can and cannot be resolved; it is resolving because B brought them coffee that morning, but it is unresolving because A cannot be certain that they will do so in perpetuity.[89] The same goes for confirming the trust; A is in a position to confirm that B brought them coffee that morning, but they cannot confirm that they will do so every morning. However, this response only highlights the inaccurate ways in which we speak of trust. The supposed contradiction only exists because there are two distinct trust instances happening, and the intuitions about the case depend on which trust type is at issue. To see this, consider MORNING COFFEE once it has been taken apart:

> DECONSTRUCTED COFFEE. A trusts B to bring them coffee that morning, but they also trust B to bring them coffee every morning.

---

[88] Keep in mind that we are not claiming that definite trust cannot exist without the trust being confirmed, but that definite trust requires a closely related act of trust confirmation that accompanies the definite trust instance, regardless of whether the trust is actually resolved or confirmed.

[89] One could argue that "every morning" might be unrealistic, and instead it should read something like "every morning as long as we are living together, the trustee is not sick or injured, not late for work, and so on.", but the point being made here holds in either case.

The former instance of trust can be resolved while the latter cannot as there is no time t at which B can bring them coffee every morning. One immediate objection that comes to mind are definite trust cases that involve conditionals where there is no opportunity for B to Φ, either because B is unable to Φ or because Φ cannot be executed. Consider the following case that challenges our definitions of confirmed trust and resolved trust:

> BENCH PRESS. Avery is at a gym, preparing to perform a bench press. Uncertain whether they can complete the lift, Avery asks a nearby gymgoer to spot them, i.e., to stay close by during the attempt and assist if the lift proves too difficult. Avery trusts the gymgoer to help them in case of failure. Fortunately, they successfully complete the lift without requiring assistance.

It looks like Avery initially affirms their trust in the gymgoer by asking for a spot, and the trust relation between them can be characterised as definite trust, as it can be resolved and confirmed. However, because Avery did not fail and thus the gymgoer had no opportunity to execute Φ (in fact no one had the opportunity to execute Φ), the trust is neither confirmed nor resolved. What is going on here? One answer is to point out that the gymgoer did in fact do as entrusted, as they fulfilled their ongoing trust obligation from the time of the trust affirmation until the lift was completed. Because there was no failure, the conditional "spot in case of failure" can simply be trivially discarded. Now, if Avery had failed the lift, then the trust conditional could not be trivially discarded, and we would end up with a standard definite trust case in which the trust would be resolved iff the gymgoer acted as entrusted, and Avery would (very much) be in a position to confirm the trust.

To recap, definite trust involves cases where the trust is limited in the sense that the trust is proceeding towards a terminus; there is an end to the trust. The set terminal point is not necessarily planned or agreed upon by both parties, but both parties are aware of the Φ and what is more, they are aware of what needs to occur for Φ to be resolved. A typical example of a definite trust case is as follows:

> WASHING MACHINE. A trusts B to hang up the clothes from the washing machine before B goes to bed, as A will be going to bed before the washing machine will complete its cycle (and before B goes to bed).

When B has done as entrusted, then A no longer needs to trust B to hang the clothes up. Furthermore, this does not necessarily require A to always trust B to hang up clothes in the future, although this sort of successful trusting often leads to A forming a stronger belief about B's trustworthiness moving forward. A and B are both aware of what B is entrusted to do, and they both know that when B has done as entrusted to do, then they have completed their part in that particular three-place definite trust.

Trust Confirmation, Therapeutic trust, and Monitoring

Recall that we said that when B has done as entrusted, then A no longer needs to trust B. This is not entirely correct. It would be unfair to blame A for trusting B to do something when B has already performed the entrusted action. In this case it would translate to A trusting B to hang up the clothes from the washing machine when there are no clothes left in the washing machine, as B would already have done as entrusted. In many definite trust cases such as WASHING MACHINE, it seems important for A to confirm their trust. When should A stop trusting B that they will hang up the clothes? Presumably after the deadline, so to speak, that was set during the trust affirmation. For A to maintain their trust in B after the deadline would be a different belief, something like "A believes that B performed the task as entrusted". If this case would not have a confirmation, then A's belief would be that they trust that B has already performed the task as entrusted. In many cases this would be an unsatisfactory ending to the trusting relation. At the very least it would be unlikely to increase A's trust in B in the long run, as A has not acquired any evidence about B having successfully performed the task as entrusted, which if obtained, could be used as evidence about B's trustworthiness in the future. This is indicative of a by-product of distinguishing between definite and indefinite trust that relates to therapeutic trust cases. For now, it suffices to say that definite trust cases can therapeutically increase trust and that therapeutic trust cases must involve definite trust to successfully build trust. This further suggests that maximally successful definite trust cases, even in non-

therapeutic trust cases, must not only be resolved, but confirmed. This makes sense when we consider the alternative, that unresolved definite trust cases, that cannot build further trust but only maintain the already established trust, are as good as definite trust cases that are resolved, which in addition to maintaining the trust are able to increase it.[90]

One might ask why indefinite trust cases cannot constitute therapeutic trust cases. Consider MORNING COFFEE once more, where it might look like A trusts B to bring them coffee every day, and when B has done so for a few days in a row A's trust in B might have increased, thus therapeutically increasing A's trust in B in an indefinite trust case. Not so fast, notice that although A has indefinite trust in B bringing them coffee, they also have the definite trust that B will bring them coffee at t_1, t_2…t_n. This aligns with what we observed in DECONSTRUCTED COFFEE. Every one of those instances is a case of definite trust that comprise the actual therapeutic trust cases responsible for the trust increase in B.

Furthermore, try to imagine what a strictly indefinite trust case would look like. In that case, A would trust B to bring them coffee every morning, and A would not need to increase their trust in B further to trust them to do so. After being handed coffee for a few days in a row, does A then have an increased trust that B will bring them coffee every morning? I find that unlikely, unless A did in fact not trust B to bring them coffee every morning, but instead trusted B to bring them coffee at times, or every day for the rest of the week. In the former case, where A trusted B to bring them coffee at times, then it is unclear whether B doing as entrusted in fact increases A's trust, as the frequency of the action has less weight than in the other cases.

In addition, it seems like A might not even trust B at all in that case, as one could argue that A cannot even rely on B bringing them coffee at any particular time, which culminates in something akin to trust without reliance. Instead one could think about this as A knowing that B brings them coffee at times without trusting B to do so. In the second case, where A trusted B to bring them coffee every day for the rest of the week, we can easily see that as an instance of definite trust that is resolved and confirmed at the end of the week, which itself is comprised of several definite trust cases that get resolved and confirmed every morning. To summarise,

[90] This depends on the notion that trust is intrinsically valuable in some way.

MORNING COFFEE shows how even when it looks like indefinite trust can be therapeutic, it is only because it entails definite trust cases that are doing the therapeutic work.

In the WASHING MACHINE example, A will eventually check whether the clothes are hanging in the washing room, or simply ask B the day after whether they remembered to hang the clothes.[91] The point is that A eventually wants to confirm whether their trusting relation was successful. This kind of *confirmation monitoring* is essential to definite trust cases and should be considered a defining feature of them. To illustrate, if A would never get any confirmation about whether B took care of the clothes in the washing machine regardless of whether the trust has been resolved, then A would have to keep trusting B to have done so indefinitely. Confirmation monitoring can be defined as follows:

> **Confirmation Monitoring:** The monitoring the trustor performs at, or after, the set terminal point of a three-place ongoing definite trust case, to confirm that the trustee has done as entrusted.

An interesting observation here is that maximally successful therapeutic trust cases are always cases of ongoing definite trust cases. If a therapeutic trust case would be indefinite, then it would not possess the necessary qualities to account for an increase in the trustee's trustworthiness. To illustrate further, we introduce two similar cases where the first one portrays indefinite trust and the second definite trust. Consider first the case involving indefinite trust:

> INDEFINITE CHARITY. A is at work and learns that there is a charity based in A's city that is doing excellent work. Inspired, A intends to donate money to the charity. Unfortunately, the only way to donate to the charity is in-person, and A is stuck at work for the remainder of the day. A is aware that their coworker, B, is going to run some business errands close to the charity's location and asks whether B would be willing to take $50 in cash and donate to the charity on their behalf. The amount of money A gives B is less than what A was willing to give to the charity but the maximum amount A trusts B with. B says he will do as entrusted and A trusts that B will do as entrusted. As

[91] It will be argued in Section 7 that this kind of scenario portrays at least *pseudomonitoring*, i.e., monitoring without intention.

> the charity does not respond to inquiries about individual donations and does not allow any kind of recording within their facility, there is no way A can know whether B really donated the money.

This is a good case of trust, but it does not portray therapeutic trust because there is nothing that indicates that A can trust B to a greater degree when B arrives back at the office without the cash. In other words, A's trust is unable to build further trust because A cannot confirm the trust by confirmation monitoring, thus lacking the main characteristic feature of therapeutic trust. Even if A trusts B and believes that B donated the money, A's trust in B is only maintained and there is no trust increase because nothing new has come to light that suggests B is more trustworthy than before.

Note that even if B donated the money, the trust would only be resolved but still unconfirmed. If one tried to argue against this, they would also have to stand by cases that do not match our intuitions on trust. For example, consider that if the relevant trusting relation between A and B was therapeutic, A might trust B with more money next time, and so on, without any confirmation of any sort that the money was really going to the charity. A's friends and relatives might ask A why they trust B with such large sums of money, and A would have to respond that they trust that B is donating the money because they believe A has always donated it in the past. The problem here is that A is in the same position they were the first time they trusted B, no new information has come to light to suggest that A should believe that B is any more trustworthy than they were the first time around. This is not to say that A would fail to form the belief that B is in fact more trustworthy if A would be in a position to confirm that B did what they were trusted to do. The point here is that there are cases where it would be epistemically wrong of A to form that belief, as they would not be following evidential norms while doing so.

Before we analyse the second case involving definite trust, it is worthwhile to explore further why indefinite trust cases cannot be therapeutic. At this point we have shown why confirmation monitoring, a key characteristic of definite trust cases that indefinite trust cases lack, is essential to increasing trust therapeutically. However, that only pertains to the trust of the trustor towards

the trustee. We have yet to discuss another way trust can be increased, namely, by the trustee becoming more trustworthy in virtue of the trust placed in them.[92]

If indefinite trust on its own can promote further trustworthiness in the trustee, then maybe there is a way for indefinite trust cases to be therapeutic after all. The idea is as follows. If A currently trusts B with $40 in INDEFINITE CHARITY, but the act of trusting B makes B more trustworthy (in virtue of having an increased sense of duty or commitment towards A for example), and A is aware of this change in B, then A might trust B with $50, making the indefinite trust in B therapeutic. This kind of trust is not therapeutic in the traditional sense, where one decides to trust someone lacking trustworthiness in order to build trust, but it is still therapeutic in another sense because the act of trusting increases the trust. The difference here can be explained in terms of the temporal concepts introduced in this chapter, namely, that traditional therapeutic trust depends on confirmation monitoring to be successful while this kind of recursive therapeutic trust can successfully increase trust from the outset of the trust affirmation.

There are two ways to respond to the claim that indefinite trust can in fact be therapeutic by recursively increasing trust. The first response is the more straightforward one in which we point out that the trustor is generally not in a position to know whether the trustee's trustworthiness has increased in tandem with the additional trust given to them. Specifically, A does not initially trust B with $50 but by taking a metaphorical leap of faith and handing over the $50 anyway, A believes that B's trustworthiness will increase, and on that basis forms the belief that B can be trusted with the extra $10. However, even though A believes that B will become more likely to do as entrusted after being granted this additional trust, they do not know whether B can be trusted without engaging in confirmation monitoring. Furthermore, it seems like A's belief, that trusting B will increase B's trustworthiness, is more akin to hope[93], or even a kind of second-order trust, because A has no evidence to support their belief; if they did, it would already be evident in their initial assessment of B's trustworthiness. For example, if A

[92] A comparison can be made here to Faulkner (2011), who suggests that a hearer trusting a speaker makes it more likely that the speaker will tell the truth.
[93] For the purposes of this chapter we can use a standard account of hope that maintains that to hope for an outcome is to desire it and to believe that it can be realised without being inevitable (Downie 1963, 248).

had evidence that B prided themselves on never failing to do as entrusted, A would be inclined to trust them more than if they did not have any such evidence.

As confirmation monitoring is impossible in INDEFINITE CHARITY, A is left with the hope that the trust relation itself has recursively increased the trust without being able to confirm that their trust has been resolved, which places A in the same situation as before. Furthermore, if A were to continue thinking that trust itself recursively increases the trustee's trustworthiness in indefinite trust cases, making them trust to a greater degree than they would otherwise, they would not only remain incapable of confirming that their trust is resolved, but they would also be unable to confirm their second-order trust that indefinite trust can increase recursively.

The second response is less obvious but nevertheless important. Imagine that B would be the kind of person who desires to be generally perceived as a trustworthy individual. However, when they see an opportunity to profit greatly, they value the profits over their reputation and conscience. B would not find $40 to be worth the guilt, but $50 would be. In this case, B becomes less trustworthy the more therapeutic trust they are given. Furthermore, even if A was aware of B's character deficit, they would still trust them with $40, as they are certain that it will not tempt B sufficiently to resort to betrayal. We now turn to the second case that portrays definite trust:

> DEFINITE CHARITY. In this example, everything is the same as in INDEFINITE CHARITY, except A receives a phone call from the charity every time A donates money to them. In this case, A trusts B with their money and subsequently receives a phone call that confirms A's trust in B. This kind of trust is therapeutic and can increase the degree of trust over time because A is in a position to confirmation monitor that their trust has been resolved.

The difference between these two cases is that the first one is indefinite, and the second one is definite from the perspective of A. Both cases have the properties of ongoing three-place trust, but only the latter can induce a therapeutic trust increase. Why is that? The sort of monitoring happening in DEFINITE CHARITY, that we have defined as confirmation monitoring, confirms that A was right to trust B.

We see that monitoring thus plays a vital role in therapeutic trust cases. Consider a paradigmatic therapeutic trust case, in which parents leave their teenager at home to watch the house while they go on a weekend holiday.[94] They trust that the teenager will not throw a party and keep things tidy while they are gone. When the parents return, they will notice either that the house is a mess or that it is tidy. Either way, they will have some evidence about whether there was a party at the house. Of course it can be argued that the teenager could have thrown a party and cleaned the house meticulously afterwards, lied to their parents expertly, bribed their neighbours not to alert the parents, etc., but this line of argument is not entirely convincing. The parents would not increase their trust in the teenager without at least believing that their teenager did not host a party, regardless of whether their belief is true.

One way to criticise this conclusion is to say that even if one normatively *should not* increase one's trust after trusting someone in indefinite trust cases, many people still do, and that is sufficient to claim that therapeutic trust can take the form of indefinite trust cases. After all, therapeutic trusting can fail but one still trusted with the aim of building trust. One way to respond to this is to examine closely how much of this criticism is rooted in the notion that trust is voluntary. However, even paradigmatic examples that purportedly show how trust can be voluntary,[95] such as Richard Holton's (1994) central case, are unconvincing:

> TRUST FALL. You are blindfolded and stand in the middle of a circle of peers. They make you spin until you lose your bearings and then, with straight legs and arms by your sides, you let yourself fall (Holton 1994, 63).

This case is supposed to show that when faced with a trust fall exercise, we will ourselves to trust. I am inclined to believe that one can consider the risk of falling as the price one has to pay in order to find out whether the people are trustworthy. We do not know whether the group

[94] This case in particular is adapted from Karen Jones (2004, 16).
[95] For further reading on voluntary trust, see e.g., Booth (2018), who states that doxastic accounts of trust cannot explain that we can at times trust at will, in a way we cannot do with regards to beliefs. He suggests this is because trust should be thought of as a term given to mental states that we would consider beliefs if belief were to be thought of as a state posited primarily to explain the actions of agents (2018, 1).

will catch us, but we decide to trust that they will in order to find out.[96] During the fall it seems obvious that one either expects to fall to the ground, which would not be trusting, or be caught before falling, which would be trusting. One can will oneself to fall, but an expectation cannot be willed in the same way. TRUST FALL is a definite trust case, and one endures the risk to get confirmation whether or not the group can be trusted by eventually monitoring for evidence (namely, whether one was caught or not). In indefinite therapeutic trust cases there is no eventual monitoring, and no confirmation that you were caught by the group. If one cannot voluntarily[97] trust, then building trust upon indefinite therapeutic trust exercises is no better than simply deciding to increase one's trust by sheer will, with no therapeutic trust involved.

To capture the distinctions made in this section, refer to the following table.

	Prior to trust affirmation	**In case of definite trust**	**In case of indefinite trust**
Precautionary measures that do not undermine trust	Evidence, monitoring, reflecting.	Confirmation monitoring.	None.
Can be resolved/therapeutic	N/A	Yes.	No.

[96] Of course, one could be certain that they will be caught given the right circumstances, but it is easy to imagine a TRUST FALL scenario in which one genuinely does not know whether they will be caught, which is the kind of case we are interested in here.

[97] For a defence of the view that belief can be voluntary, such as the belief that someone is trustworthy, see literature on doxastic voluntarism, for example, Ginet (2001).

The Risk of Betrayal and the Risk of Disutility

We need to make one more distinction. Within the epistemology of trust, we find that the risk of betrayal is generally used to signal the harm of feeling betrayed.[98] However, when we examine precautionary measures, we find that they oftentimes have limited or no mitigating effect on the sort of harm that accompanies betrayal. Instead, these precautionary measures are taken to mitigate the harm of the trustee not doing as entrusted.

For further clarification, consider that trust relations come with their own moral norms (Dormandy 2019, 5). It is wrong to steal a cake (harm of disutility), but if the thief has been trusted not to steal it, and they still steal it, they have committed an additional moral infraction by betraying the trust placed in them (Dormandy 2019, 5). On one hand, we can talk about the risks of being let down or feeling betrayed. On the other, we have the risk of being harmed by the trustee's failure to execute the action they have been entrusted with. We can call this *the risk of disutility*. In other words, there is the risk of the trustee not Φ'ing (risk of betrayal), and there is the risk of ¬Φ (risk of disutility).

Three-place trusting always carries the risk of being let down, i.e., the risk of the trustee not Φ'ing, and this is true for both definite and indefinite cases. This risk of being let down can be characterised as the risk of the betrayal itself, regardless of the actual impact of ¬Φ. We can define this as the risk of being disrespected, although the harm often takes the form of experiencing a feeling of betrayal.[99] This can include the realisation that your trust was misplaced, that your trusted friend might not value your friendship, or simply a general feeling of unfairness. Note here that betrayal on its own, even when there is no other disutility at stake, wrongs the betrayed. For clarification, compare betrayal to lying. When someone lies to you, they are disrespecting you in some way even if you do not, for any number of reasons, experience it as a lie.

[98] See e.g. Holton (1994), Dormandy (2019), and Nguyen (2022).
[99] For example, some philosophers who endorse responsiveness theories of trust, such as Jones (2012) and Faulkner (2007), find betrayal to be "grounded in the betrayer's failure to be properly responsive" (Nguyen, Trust as an unquestioning attitude 2022, 216).

Let us now turn our attention to the risk of disutility. Interestingly, we find that in some cases of definite trust, there is little to no risk of disutility, i.e., the risk of Φ not being executed, although the risk of being let down would still materialise if the trustee would not be the one to execute Φ. When trusting someone, it depends on the Φ at hand whether the definite trusting can create the risk of ¬Φ.

Consider the following two examples. In the first, A trusts B to water the plants while A goes hiking the Appalachian trail without any electronics (to be closer to nature). In this case, we see that A trusts B and then maintains ongoing definite trust until they get back to the house to see their plants either alive or dead. Whatever B decides to do, the plants would be just as alive (or dead) regardless of A's actions. There is no way for A to monitor B, and thus no way to minimise the risk of ¬Φ. In the second example, consider that A had their phone with them, and been notified shortly after leaving that B had gone on a spontaneous holiday abroad, having completely forgotten their plant-watering duties. In that case A could make the necessary arrangements for someone else to water the plants, thus avoiding the risk of ¬Φ, but A would still have to deal with the materialised risk of being let down. This distinction is thus significant to the general aim of trust-related precautionary measures, such as monitoring, because it illustrates how they are not capable of eliminating all of the risk that trust entails. Furthermore, this distinction enables us to argue later in this chapter that it is not monitoring that undermines trust, but the intention to do so.

To further motivate this distinction we can call on the philosophical literature on consent, as trust and consent are analogous in some ways, and interconnected.[100] For one, consent and trust both change which norms are in play; consent eliminates rights and duties (Dougherty 2021, 56), while trust creates them. They can both be embedded in complex moral agreements, as one can decide to consent *because* they trust (Dougherty 2021, 56). Furthermore, it is easy to see how one can both signal consent and trust in a single action, when I trust someone to feed my cat I am also consenting to them feeding my cat (Dougherty 2021, 56). On a similar note, if I

[100] I am repurposing Dougherty's ideas on the connection between promises and consent, as promises are oftentimes cases of trust. To read more about the connection between promises and reliance, see MacCormick (1972), and for further discussion on the relation between promises, trust, and cooperation, see Fried (1981) and Prichard (2002).

revoke my consent that they can feed my cat, I also stop trusting them to feed my cat. (Dougherty 2021, 56).

To better grasp how one can incur harm in cases of betrayal even when no risk of disutility materialised, we can examine how nonconsensual sex is by itself sufficient for moral wrongness, even though other harms can then make the wrongness even greater. One way to defend the thesis that nonconsensual sex is seriously morally wrong is to fall back on a view proposed by Alan Wertheimer, which roughly states that the seriousness of the wrong of having sex with an unwilling unconscious person can be explained by the harm suffered by the victim (Dougherty 2013, 725). Tom Dougherty calls this *the harm explanation* and proceeds to argue that it fails to capture the wrongness of having sex with an unconscious person because there is not necessarily harm involved.

Firstly, the sex itself may not be physically harmful, secondly, as the victim is unconscious, they do not suffer experiential harm, and finally, if the crime remains undetected the victim will not suffer psychological harms at a later stage (Dougherty 2013, 725). This seems obviously wrong and does not match our intuitions about cases like these. To defend the thesis that nonconsensual sex of this sort is seriously wrong, Dougherty instead proposes that the seriousness of the wrong of having sex with an unconscious person is explained by the fact that the victim did not validly consent to the sex (Dougherty 2013, 724-725). This view is compatible with the notion that "[h]arm makes an action worse, even though its nonconsensuality is itself sufficient for the action's wrongness" (Dougherty 2013, 727). Dougherty frames this nicely by writing: "If a stranger trespasses in your garden, then her action is wrong in virtue of the fact that she lacks your consent. But it is worse if she thereby ruins the flower beds" (Dougherty 2013, 727).

This notion of nonconsensuality being by itself sufficient for moral wrongness, even though other harms can then exacerbate the wrongness, can be straightforwardly applied to trust. In much the same manner as was previously argued, one can imagine a trust case in which Taylor trusts Drew to take care of the flower beds and Drew fails to do so. In this example, Taylor's risk of betrayal materialises, which not only causes Taylor harm in the form of ruined flower beds, but also harm in virtue of them being betrayed and thus disrespected (which can entail

experiencing negative emotions in light of the betrayal). To further support this, consider if Drew would have failed to take care of the flower beds, but the flowers, by some miracle, stayed healthy. In that case, Taylor would still feel like they had incurred some harm, and the understanding of the trusting relation between Taylor and Drew made Taylor aware of both kinds of risk when they trusted Drew to take care of the flower beds.

One could object by pointing out that the harm that occurred in the latter case is not related to trust at all, rather, it is the harm of discovering that Drew is untrustworthy, or even the harm of Taylor realising that their sense of who is trustworthy is flawed. However, objections of this sort do not pose serious threat when we further examine the harms. Namely, the harm of discovering a fact is not inherently harmful, rather, the harm relates to the way things are, regardless of the information state of those harms. To illustrate, consider that the speech act of a doctor telling a patient that they have a terminal illness is not harmful. If anything, it can be immensely helpful for various reasons, e.g., the patient finally knows what has been ailing them, they can start preparing their future with more certainty, and they have a better idea of what changes they can make to improve their quality of life. The harm is the illness itself, not the knowledge pertaining to it. In the trust scenario, Taylor experiences negative emotions towards Drew, but those emotions are distinct from the emotion of feeling betrayed. To see why, consider that one can feel betrayed even when there are no negative emotions towards the trustee nor towards oneself. See the following Gettierised trust case:

> BIRTHDAY PARTY. Ridley trusts Robin to Φ, namely, organise Ridley's upcoming birthday celebration, as they have been friends for years and are a part of the same social circle. Among other things, Ridley trusts Robin to find a venue, decorate it, invite guests, and so on. When Ridley arrives at the venue with one of her dear friends, Wiley, all their friends greet them with birthday wishes and confetti. The venue is perfect for the occasion, and it is beautifully decorated. Later in the evening, Robin confesses that they misunderstood Ridley, and they thought they had been entrusted with Φ*, namely, planning a surprise birthday party for Wiley. Unbeknownst to Robin, Wiley's birthday is months away, and none the wiser that the celebration was apparently planned with them in mind.

In BIRTHDAY PARTY, Ridley experiences no harm because of Robin's betrayal (by not performing Φ but Φ*), and even though the betrayal is not egregious[101], and Robin meant well, one can be sympathetic to Ridley feeling like they have been betrayed, and that they might find their happiness to have decreased somewhat after discovering the miscommunication. Note that our response to the previous objection, that the harm of betrayal relates to the state of things rather than discovering the state of things, does not have as sharp of an edge in this case, because Ridley wants the state of things to be exactly as they are, the birthday party is perfect as is; the only thing missing is some intrinsic value in the intent of the trustee as they planned the party. A lack of this intrinsic value of trust is in and of itself sufficient to cause harm, regardless of consequences or the state of things.

To recap, there are two distinct risks involved in trusting: the risk of being let down, and the risk of ¬Φ. We have explored multiple cases involving trust where monitoring can only partially guard against the risk of being let down and without having an effect on the risk of ¬Φ. Imagine the following case:

> CAR. You lend a friend your car while you are away on holiday. You trust them to take care of it while they have it, drive safely, and make sure that they let you know if anything goes wrong. Although you initially affirmed your trust by lending them your car, you start reflecting on your trust and come to the realisation that you should not have trusted them. This reflection decreases the trust you have in your friend, so much so, that you decide to monitor them by calling them, and search for evidence by calling your neighbour to ask whether they have seen the car around and if so, whether they happened to spot the condition of the car. Now, at this stage, the trust you have in your friend is only a fraction of what it was when the trust was initially affirmed.

Relating to the risk minimising effects of precautionary actions, we see that even if you were to perform a precautionary action such as monitoring by calling your friend, you are not shielded from the risk of being let down. If your friend would then tell you that they were

[101] Hinchman might consider this a situation in which Robin betrayed Ridley's promissory trust, but not in a way that calls for reactive-attitudinal response (Hinchman 2021, 93).

driving after drinking heavily, it would be difficult to claim that your friend has not let you down. With regards to the second risk, the risk of ¬Φ, we find that regardless of what precautionary actions you take, you can only become aware of whether Φ or ¬Φ, you cannot make it so Φ. In other words, if the car were to be damaged, causing a substantial financial loss to you, it would be damaged regardless of the precautionary actions you could take, and your wallet would still suffer. Note that this is not universally true. In some cases you could get insurance or have another friend intervene, but in others you might have forgotten your phone and have no precautionary actions available at all.

Precautionary measures differ between the two kinds of risk, both with regards to how they affect the trust, and the way they minimise the two kinds of risk involved. To reiterate, the risks are the risk of betrayal, which can be characterised as the harm of being disrespected and the potential negative experience associated with betrayal, and the risk of disutility, which is the harm that results from the entrusted action not being executed.

We can conclude the following; firstly, the only way to defend against the risk of being let down, is by trusting well, for example by gathering evidence and reflecting on it, and even monitoring the potential trustee before trusting them.[102] Secondly, that precautionary measures in ongoing trust cases cannot guard against the risk of being let down and what is more concerning, can only partially guard against the risk of disutility involved in ¬Φ. They can only partially guard against the latter, as while they may reveal to the trustor that the trustee will not do as entrusted, allowing the trustor to take precautionary actions (such as buying car insurance), that is not always the case. In some cases it is impossible to intervene in order to mitigate or eliminate the repercussions of ¬Φ.

Showing how precautionary measures have generally been thought to undermine trust at least to a degree[103] is necessary to show how temporal elements of trust, such as confirmation

[102] This does not include therapeutic trust cases of the sort Arnon Keren (2014) introduces, i.e., cases where there is genuinely no trust in place, but it does include the limited-trust cases found widely in the literature, such as where parents trust their teenager to look after the house in order to promote trustworthiness.

[103] For example, Carter (2024) finds that monitoring is incompatible with trusting to the degree that, through monitoring, one aims to render oneself invulnerable to all but *de minimis* risks of betrayal one is subjected to in virtue of trusting (2024, 133-134).

monitoring, can alter the way we think about trust and precautionary measures. Some mild precautionary measures do not necessarily lead to the trust relation to break down completely. Other, more egregious, precautionary measures completely negate either the significance of the trusted action or either kind of risk involved (betrayal or disutility), resulting in complete trust breakdown. For an example of the former, imagine the following case:

> AILING DOG. Harper would not trust anyone, not even their most trusted friends, to take care of their ailing dog, that requires strict adherence to a meticulous medical regimen, without some sort of supervision. Furthermore, they would not trust acquaintances or coworkers, even with supervision. Going away for a few days, Harper proceeds to trust their close friend to take care of their dog with supervision.

The close friend in this case would still feel a sense of being trusted even when being monitored. This suggests that monitoring can bridge the gap between the degree of trust the trustor has in the trustee, and the greater degree of trust the trustor requires to trust the trustee for a task. If the close friend believes themselves to be more trustworthy than the trustor believes them to be, they will feel like they are not being trusted to the degree they should be.

One could criticise this example by pointing out that there is something crucially different between this kind of monitoring and the more typical cases of monitoring, like in Wanderer and Townsend's (2013) nanny-cam case, where parents monitor their babysitter using a hidden home security camera. I argue that the difference between the two examples has not been sufficiently addressed. Namely, that in the former, the friend that is taking care of the dog is aware of the monitoring taking place, call this disclosed monitoring:

> **Disclosed monitoring:** A takes precautionary measures when trusting B to Φ, such as reflecting on their beliefs, monitoring B, or searching for evidence whether B will Φ, and B is aware of A's precautionary measures.

In the latter case, the babysitter is being subjected to monitoring they are not aware of, call this undisclosed monitoring:

> **Undisclosed monitoring:** A takes precautionary measures when trusting B to Φ, such as reflecting on their beliefs, monitoring B, or searching for evidence whether B will Φ, and B is unaware of A's precautionary measures.

Does it matter whether the type of monitoring in AILING DOG is disclosed or undisclosed? Not necessarily, we can see that if the friend took themselves to be trustworthy up to the point of taking care of the dog *with* supervision, they might find A to be foolish trusting them to do so without supervision. Disregarding the ethical worries of the case, if it were to be revealed afterwards that they were being monitored (and the friend thought everything had gone well), they might feel relieved getting confirmation that they did a good job. Now, if the friend deems themselves trustworthy enough to take care of the dog without supervision, they might feel disappointed upon discovering that they were being monitored. However, this would not result in a termination of their trust. In contrast, if Harper would ask someone else to take care of the dog instead, then that would indicate a complete lack of trust. Consider here that even when the close friend is being monitored, they are still in a position to take care of the dog, and if they fail the consequences are dire. The standard precautionary measures, such as monitoring, reflecting, and searching for evidence, suggest that there are some doubts about the trustee's trustworthiness, while the second kind of precaution depicts a complete lack of trust in the trustee. If Harper would, instead of monitoring their close friend, ask a different friend to take care of the dog, then there would be no trusting action available to the close friend. In the former case the close friend could at least prove themselves and increase Harper's trust in them therapeutically. In the latter case that option has been removed.

It is enough for us to show how monitoring and other precautionary measures are generally thought to undermine trust at least to a degree. However, one could argue that precautionary measures can lead to a complete trust breakdown by eliminating the need for trust by either completely negating the significance of the trusted action or the risk of disutility.

Adam Carter (2024) claims that the existence of mitigating back-up plans does not preclude cases from having been cases of therapeutic trust cases because the vulnerability to betrayal is not eliminated by taking steps to mitigate damages if the risk were to materialise (Carter 2024, 143, 2022, 41). Returning to the CAR case, if you had insurance that would cover the material

cost of the car in case of a crash along with covering any inconveniences caused by no longer being able to access a car and so on, it seems odd at first to agree to his claim. After all, if we imagine that the insurance payout to be ten times higher than the value of the car, it raises the question what exactly we are trusting our friend with. A more direct example would be to have your friend provide you with a monetary amount that matches the value of the car, which they would then receive back once they return the car. In that case it seems obvious that no trust *needs* to be involved, it is simply a transactional agreement. However, when we consider the two risks of trust, the risk of being let down and the risk of ¬Φ, we find that Carter's claim is correct, as one cannot insure against the experience of betrayal (i.e., there are no mitigating back-up plans that can adequately cover betrayal).

Finally, even if we would concede that even disclosed monitoring cannot coexist with trust, there is a different response readily available. Because AILING DOG is an example of an ongoing definite trust case, there will always be confirmation monitoring when Harper, the dog owner, arrives back home and can plainly see whether their dog is healthy or not. So the issue the close friend takes with being monitored is not really about the monitoring itself, as they would happily accept the terms of the trust affirmation of taking care of the dog so that the dog is healthy when the owner arrives back. The crux of the problem they have with the undisclosed monitoring has to do with Harper lying to them (or at the very least withholding relevant information in a way that could be considered a "white lie").

This idea of precautionary measures being less effective than generally thought is relevant because it illustrates the significance of evidence gathering at the onset of trust, as that kind of precaution is most effective at guarding against betrayal without jeopardising the trusting relation that can be formed thereafter. One could object and say that it is possible to trust on a whim, for example when you are doing a pressing task that suddenly requires you to trust someone with something fast. However, this would be akin to hope, as you would not feel betrayed if the trustee did not do what you wanted them to, although you might feel disappointed in yourself for creating such circumstances to begin with. For a very different example, consider a scenario in which you are stranded on a remote road in the middle of the night, and you have no evidence to support the belief that someone will drive past and help you out. No one drives past you that night, but it would be odd to claim that you trusted that someone

would show up, and even more odd, that you feel betrayed. These cases are different as one relies on low probability while the other relies on low time, but both cases involve a lack of evidence for why one should trust. Going by a doxastic account, trust requires a belief, and the standards of belief are not dictated by practicality.

There is a rift between precautionary measures that are supposed to mitigate the risk of trusting, and the undermining of trust. If we are to understand trust undermining in relation to the mitigation of risk when trusting, then seeing how these sorts of risk minimising behaviour are less effective at minimising risk than they are made out to be should result in the trust undermining to be of a lesser degree than has been generally agreed upon. However, this result is lacking, as when we compare it against our real-world experiences, it does not hold up. If risk mitigating actions are less effective than they have been made out to be, why do they still undermine trust to such a large degree?

One response is that the intention[104] to mitigate risk at the first order without thinking about trust as such, is what causes the undermining, rather than the subsequent actions such as monitoring. However, this kind of a response would be problematic, as it allows for cases in which someone mitigates risk for reasons other than a lack of trust while still causing the trust to be diminished. For example, consider the case of Taylor and Drew once more, in which Taylor trusts Drew to take care of the flower beds. Taylor trusts Drew to water the flower beds, and Drew waters the flower beds as entrusted. However, a few years earlier, Taylor had offhandedly asked their green-fingered neighbour if they were willing to water the flower beds in case he forgot. As it turns out, Taylor's neighbour is exceptionally diligent, and has kept a watchful eye on Taylor's flowers all those years with their watering can ready at hand. However, because Taylor always remembered to water the flower beds the neighbour never had to intervene. Now, does Taylor's agreement with their neighbour diminish Taylor's trust in Drew? If such intentions to mitigate risk at the first order would affect trust, we would be forced to concede that Taylor's trust in Drew is diminished.

[104] Section 6 contains further discussion on intentions and precautionary measures.

For a quick summary we have found that trust can either be definite or indefinite, and that for definite trust cases to be considered maximally successful they must eventually be confirmed (which entails that they must be resolved). This attribute cannot be found in indefinite trust cases, leading us to claim that therapeutic trust cases must necessarily be definite trust cases. Furthermore, we demonstrated that monitoring is essential to confirm definite trust cases. Finally, we showed that precautionary measures can only account for one kind of risk associated with trust, while the risk of being betrayed cannot be mitigated while in a state of trust. We are now in a position to apply our findings to three different views on trust. First, we will show that Arnon Keren's (2014) preemptive reasons account is mistaken because there can be trust without responding to preemptive reasons. Then, we will introduce some challenges for Emma Gordon's (2022) view on how monitoring can facilitate trust. Finally, we will explore in more detail how to alleviate the tension between trust and precautionary actions, by criticising Wanderer and Townsend's (2013) proposal on how to solve the tension between trust and rationality.

Preemptive reasons

What impact, if any, do the distinctions presented in this chapter have for views about the nature of trust? Whether intentional or not, some trust accounts fail to recognise how temporal elements can alter trust cases. This leads to confusing situations where an account of trust is perceived as being capable of capturing a wide range of trust cases because it can explain a paradigmatic example of trust, when in fact it is limited to a specific kind of temporal trust and cannot account for other paradigmatic trust cases that possess different temporal elements. When the subtle differences between these cases are accentuated, it looks like not all accounts are as suited to explain them as initially intended. One such account is Arnon Keren's (2014) doxastic account of trust that involves preemptive reasons.

Keren offers an account that states that reasons for trusting are second-order preemptive reasons to not engage in precautionary actions such as monitoring, rational reflection, or searching for evidence. We can succinctly summarise his view as follows. When someone invites you to trust them, they are giving you reasons to trust them, and those reasons for trust also serve as

additional second-order reasons why you should not doubt their trust by monitoring, searching for counterevidence, or reflect on the trust. This account of preemptive reasons can be paired with either a doxastic or a non-doxastic account of trust. Keren opts for a doxastic account of trust,[105] and it seems that a doxastic account coupled with preemptive reasons can withstand the problems doxastic accounts usually face, namely, that trust seems to be different from beliefs in some critical areas. For example, while searching for evidence and reflecting on said evidence can bolster a belief, the same practice seems to definitively undermine trust. Keren's account of trust can be stated as follows:

> **Keren's doxastic preemptive reasons account:** A trusts B to Φ only if A believes that B is trustworthy, such that in virtue of A's belief about B′s trustworthiness, A sees themselves as having reason to rely on B′s Φ'ing without taking precautions against the possibility that B will not Φ, and only if A indeed acts on, or is responsive to, reasons against taking precautions (Keren 2014, 27-28).

Keren's motivation is to introduce an account that can explain how trust is not incompatible with beliefs[106]. If one rationally reflects on a belief that is supported by evidence, the belief will generally not be undermined but made stronger. This is typically not the case with beliefs involving trust, as any attempt to eliminate the risk of trusting by engaging in precautionary activities, such as reflecting on the evidence for and against the trustee's trustworthiness or monitoring the trustee to see whether they are in fact trustworthy, tends to undermine said trust (Keren 2019, 114).

The specific claims of Keren's that are of relevance here are that "[…] reasons for trust, quite generally, involve second-order reasons against acting for precautionary reasons" (Keren 2014, 15), "[…] unless we are responding to preemptive reasons, we are not trusting" (Keren 2014,

[105] Doxastic accounts of trust can explain how trust can give reasons for belief, which is challenging for non-doxastic accounts of trust because they need to rely on something non-doxastic to explain how it is possible to form beliefs from trust. One way this challenge presents itself is that proponents of non-doxastic accounts of trust generally agree that it is impossible to trust someone to Φ while believing that they will fail to Φ, or that they are not trustworthy with respect to Φ'ing, but they have a hard time elucidating this impossibility. For further reading on the differences between doxastic and non-doxastic accounts of trust, see e.g., Holton (1994), Faulkner (2007), Frost-Arnold (2014), and Keren (2019).

[106] For a discussion on preemptive reasons for belief, see Zagzebski (2012).

21), and "[f]or in trusting we see ourselves as having preemptive reasons against taking precautions" (Keren 2014, 14).

I would like to push back against these claims and argue that there are multiple cases of definite trust where we can trust without responding to preemptive reasons. Furthermore, I argue that even if preemptive reasons are not present in many paradigmatic cases of trust, they are still a valuable addition to the landscape of trust, and still a necessary factor in explaining indefinite trust cases.

When we consider definite trust cases, we find that reasons for trust are not second-order preemptive reasons simply because there are no precautionary measures that can be taken until after the trust affirmation has been established. The only precautionary actions that can be identified before and during the trust affirmation are actions that aim to analyse whether there is a good reason to believe the trustee to be trustworthy, and because no trusting relation has been established, these precautionary measures do not diminish the definite trust that follows the trust affirmation. In a definite trust case the trustor can take precautionary measures before affirming their trust in the trustee. These precautionary measures would be considered preemptive reasons if the trustor already trusted the trustee, namely, reflecting on their evidence regarding the trustee's trustworthiness, monitor whether the trustee makes good on his promises to others, and search for evidence regarding the trustee's past trusting relations. Then, if A comes to believe that B is trustworthy, then A trusts B by affirming their trust in B, at which point there are generally no immediate preemptive reasons available to respond to at that stage of the trusting relation.

After the trust affirmation, we find that the reasons why A trusted B are also higher order preemptive reasons for not partaking in precautionary measures. However, even if the reasons for trust that initially induced the trust affirmation carry with them preemptive reasons into the ongoing trust, it does not mean that the initial reasons for trust act as preemptive reasons during the trust affirmation. To illustrate, preemptive reasons are reasons for not taking precautionary measures. If there are no precautionary measures to take, then what becomes of the preemptive reasons? In other words, what exactly does the response to preemptive reasons entail, and should this response be considered an action or an intention?

In ongoing definite trust cases we found that some kind of confirmation monitoring is necessary for the trusting relation to maximally succeed. Unlike the act of trust affirmation, where one can monitor outside of trusting (before the trust begins) this confirmation monitoring cannot happen without ongoing trust present, as the ongoing trust will not be confirmed until the confirmation monitoring occurs. As an example, consider WASHING MACHINE once more. In that case, A trusts B to hang up the clothes from the washing machine before B goes to bed. Although this definite trust is resolved when B has taken care of the clothes, it is not confirmed until A has gathered evidence that B has done so. For A to be able to therapeutically increase their trust in B, they must eventually monitor the task B was entrusted with for confirmation. If A would never be able to confirm whether B did as entrusted, then the trust would not be completely successful. To see why, reflect on the fact that without confirmation, A would be forced to continue trusting B indefinitely, making this an indefinite trust case. We have previously concluded that therapeutic trust cases cannot be indefinite, which would mean that WASHING MACHINE would be a case of trust that could have had therapeutic value but because of a lack of confirmation the opportunity for a therapeutic trust increase is wasted.

This kind of confirmation monitoring is incompatible with a doxastic trust account that involves preemptive reasons because ongoing definite trust cases require one to eventually stop responding to preemptive reasons to confirm that the trust has been resolved. To clarify, in definite trust cases, the trustor trusts the trustee until the trustor has confirmed through monitoring whether or not the trustee has done as entrusted. Naturally, at the time of the confirmation monitoring the trustor is still in a state of trusting. If they were not, then they would be unable to monitor whether the trust has been resolved because there is no trust at hand to resolve.

Although preemptive reasons are either not present, like during trust affirmations, or must necessarily be ignored, like in ongoing definite trust cases, a preemptive reasons account such as the one Keren presents seems integral to ongoing indefinite trust cases. These cases depict ongoing trusting scenarios without final resolutions in which no confirmation monitoring is required, and it is evident that monitoring, searching for evidence, or excessive reflecting, undermine the trusting relation.

This is an unsurprising result. Although Keren extrapolates his notion of preemptive reasons from cases of speaker-trust, which can generally be categorised as trust affirmations (e.g. A trusts B to be truthful when they testify that *p*), it is possible to see speaker-trust as ongoing indefinite trust cases by broadening our view. Consider that when B has testified to A that *p*, A must then continue to trust in B's testimony while harbouring the belief that *p* until an unspecified time (for example if A ignored preemptive reasons and gathered first-hand evidence about *p*, thus no longer needing to trust in B's testimony). One might also argue that speaker-trust instances can even be thought of as ongoing definite trust cases by claiming that the trust takes the form: "A trusts that B *will* be truthful when they testify that *p*", in which case the definite trust would be resolved when B testifies truthfully that *p* to A. This line of thinking would be misguided however, as although speaker-trust plays a part in the entrusted action in this specific three-place definite trust scenario, it is not a case of speaker-trust per se.

We can assume speaker-trust to be a form of trust affirmation by the hearer. This is further supported by drawing attention to the difficulty of trying to ascribe preemptive reasons to speaker-trust, as there are no precautionary measures A can take during the trust affirmation, which would be a trivial task if speaker-trust were a subset of either definite or indefinite trust. Furthermore, if we argue that it is possible to view this as an ongoing definite trust case, in which A trusts that "B *was* truthful when they testified that *p*", it becomes unclear how A can respond to preemptive reasons, as A is trusting that something was the case in the past. In addition, consider that if A only meant to trust B that *p* until they confirmed that *p* by looking at other evidence, B would not feel like A had trusted them that *p*.

More generally, speaker-trust takes the form of a trusting relation where A believes that B is trustworthy (and those that are trustworthy can be trusted with telling the truth). When comparing the two examples Keren gives about speaker-trust to argue for a preemptive reasons account, one of the cases involves a speaker asking the hearer to take their word for it that *p*, while the other involves a speaker that provides good evidence that *p* (but does not ask for trust and does not give any preemptive reasons to the hearer) (Keren 2014, 11). Keren's idea here is that in the former case, the speaker would not criticise the hearer for ignoring other evidence, while the speaker in the latter case would be in their right to criticise the hearer for not taking other evidence into consideration (Keren 2014, 12). According to these cases, both testimonies

count as evidence while only the former gives reasons for trust that operate as second-order preemptive reasons not to search for other evidence.

Although these cases are convincing there is one thing that must be addressed. Specifically, that both cases exhibit trust to some degree. The former case is a typical case of trust affirmation, where the hearer affirms their trust in the speaker by taking their word for it that p. In the latter case however, when the speaker gives good evidence that p, they are also providing reasons for why they should be trusted by the act of testifying. It is not far-fetched to think that the hearer, after hearing the speaker's testimony in the latter case, would form the belief that they trust that the speaker would not provide them with good evidence that *p* while concealing that they have stronger evidence that not p. Consider the following case: A doctor provides a patient with compelling evidence that they have a terminal illness. The patient trusts their doctor and forms the belief that they have been diagnosed with a terminal illness. It seems strange that the doctor would then say that they should have looked at other evidence before forming that belief. Even stranger, the doctor could possess greater evidence that the patient does not have a terminal illness that they decided to withhold from the patient. The testimonial act of the speaker is in and of itself evidence as well; the fact that the speaker decided to present good evidence that *p* must be taken into the account.

Responding to preemptive reasons seems to be a cornerstone of maintaining ongoing indefinite trust. We can still challenge this, consider the following example:

> PHONE COUPLE. A couple, A and B, are in a healthy trusting relationship; they have invited each other to trust that they will be faithful in their relationship and given preemptive reasons for that trust. However, A really wants to check B's phone but refrains from it due to those preemptive reasons. Meanwhile, B does not even think about monitoring, reflecting, or searching for evidence regarding A's fidelity. They have simply never thought about it and are unaware that they are responding to any preemptive reasons; they simply trust the other person.

Keren claims that trust is not compatible with seeing oneself as having no reason against taking precaution and that "[t]he extent of our trust is at least partially determined by the degree to

which we see ourselves as having a preemptive reason against taking precautions and the degree to which we act accordingly" (Keren 2014, 20). It seems like B would not see themselves as having no reason against taking precaution, and in fact B does not see themselves as having any reasons for or against precaution. They just do not think about precaution at all.

This case shows that when A reluctantly decides to not look for evidence, they are responding to preemptive reasons. It seems counterintuitive that Keren would want to argue that A exhibits a more optimal form of trusting than B, as A is actively responding to preemptive reasons after reflecting on their trust, because such reflection is thought to undermine trust. As Nguyen has pointed out, it seems like "[t]rust only comes to mind once it has been threatened" (Nguyen, Trust as an unquestioning attitude 2022, 15). Simply thinking about a trusting relation, and the precautionary measures one could take, indicates that the trusting relation is already on shaky grounds. At the very least it rests on a more fragile foundation than if one did not have the desire to analyse the trust. This means that trust can be undermined by reflection before any monitoring has occurred.

The common view here is that the reflection on its own has the potential to undermine trust, regardless of what the reflection brings about.[107] I propose that it is the intention to act out precautionary measures that undermine trust, not the precautionary actions themselves. To further support that claim, imagine if someone who trusted you revealed, after you had successfully done as entrusted, that they really wanted to monitor you but decided not to do so. Now suppose that they had actually monitored you by mistake, as they intended to monitor someone else. If one had to choose between those two cases, it looks like it would be preferable to be the one that is being (accidentally) monitored.

One objection to PHONE COUPLE and other such ongoing indefinite trust cases, is to say that responding to preemptive reasons is a figure of speech, and this response can be completely passive, especially in ongoing indefinite trust cases. For clarification, imagine that B is responding to preemptive reasons unconsciously, that is, they do not think about monitoring as a possibility, and if they were made aware of the possibility of monitoring, their preemptive

[107] See for example Baier (1986) and Faulkner (2007).

reasons would still be relevant, and they would not dismiss them. Even if that were true, it still does not explain the difference between A and B in the example as they would both be responding to preemptive reasons, where responding to preemptive reasons can be said to be having the right kind of reasons against taking precaution. This would mean that actively responding to preemptive reasons would indicate a lesser degree of trust than passively responding to them, as it is challenging to accept that A and B exhibit the same degree of trusting, and if that is not the case, it renders preemptive reasons invisible in the sense that it would be difficult to consider them to be reasons to act at all.

Another way to object here is to say that B is *blindly* trusting, which can be characterised as a kind of irrational trust, where they would continue to trust even when they should not, while A trusts in a manner consistent with what we think good trust should look like. For further clarification, although B is not performing any precautionary measures, they might accidentally find strong evidence that suggests A is unfaithful, evidence that precautionary measures such as checking A's phone would also find, but because B is blindly trusting, their trust in A remains unaffected. However, there is a distinction to be made between fully blind trust, i.e., blindly trusting without reasons, and blind trust as a result of a non-blind trust affirmation, i.e., the trust was initially affirmed because there were good reasons to trust. In PHONE COUPLE, it stands to reason that B has many reasons to believe in A's trustworthiness, and can thus show exemplary trust simply by continuing placing their trust in A. This further strengthens our case, because this sort of blind indefinite trusting would oftentimes be considered admirable as long as the trust affirmation was rational. In short, blindly affirming one's trust is problematic, but blindly trusting someone in ongoing trust cases is not.

A third objection is to claim that A's inner turmoil indicates that they do not trust B and are instead just acting like they trust B.[108] To respond we must once more look at the relationship between trust and risk. Keren notes that considering relevant evidence is one of the essential precautions one can take to guard against the risk of being let down, and other actions such as checking the partner's phone serves as precaution because it can provide evidence about the

[108] Frost-Arnold (2014) writes that the demand that trust should be distinguished from mere pretence of trust is a sensible one (13). Many authors worry that trusting without belief is too similar to acting as if one trusts without actually trusting, see e.g. Baker (1987), Holton (1994), Hardin (2002), and Hieronymi (2008).

risk of being let down (Keren 2014, 22-23). So, when A decides not to check B's phone they are acting in a way that does not produce evidence about the risk of being let down, they are not attuned to such evidence, and they remain resistant to counter-evidence by not re-examining their grounds for trust (Keren 2014, 23). So by Keren's account, A is opening themselves up to the risk of being let down by not partaking in any precautionary measures, i.e., they are letting their guard down. If B would break A's trust, A might say something like: "I knew I should have ignored those preemptive reasons", but it would not soften the blow, and A would feel betrayed because B would have let them down.

A plausible objection here is that PHONE COUPLE is not a case of genuine trust to begin with because A is only *acting* like they trust B. They might press on and point out that A could very well be opening themselves up to risk for other reasons than trust, and although reasons for trust are preemptive reasons it is not the case that preemptive reasons are necessarily reasons for trust (Keren 2014, 14). Note that Keren states that someone must be responding to preemptive reasons for the *right* reasons, i.e., because of trust, as the preemptive reasons are the same reasons as reasons for trusting, and not any other reasons such as negligence. However, trust itself is valuable. Knowing that someone monitored you can eradicate trust even if nothing came out of it. In normal trust cases where the trustee thinks they are trusted, they might still believe that the reason why they will not be monitored is not because of preemptive reasons but because in this trusting relationship they believe the trustor to value their trusting relationship enough to not compromise it by monitoring. Therefore, there are cases where a trustor might say: "I wanted to monitor you, but I refrained from doing so", without incurring negative reactions from the trustee. That is, regardless of content, just the act of refraining from monitoring signals that one's trusting relationship is valued, even if the desire to monitor is there.[109]

We can further argue that not responding to preemptive reasons that are derived from reasons other than trust (so not the *right* reasons, according to Keren) can also undermine trust even

[109] Note that there are conflicting intentions at play; one is the intention to minimise risk while the other is to keep a trusting relationship

when one is responding to preemptive reasons that result from reasons of trust. Here is a case borrowed from Keren:

> CCTV. A supervisor is monitoring an employee through the CCTV system, not because they do not trust the employee, but because they are thinking about using the footage for future training videos (Keren 2014, 20-21).

Keren claims that this act of monitoring does not classify as a dismissal of preemptive reasons, as the intention is not to monitor whether the employee is trustworthy. In fact, the supervisor trusts the employee perfectly well. The supervisor's behaviour is not meant to serve as a precautionary measure and the act of looking at the CCTV footage therefore does not undermine the trust between the supervisor and the employee. As Keren writes: "[a]cting in certain ways amounts to taking a precaution only when the action is performed for a certain reason, namely to guard against a certain risk" (Keren 2014, 21). However, this can be criticised in the following way. People in general are aware of the typical precautions one could take in cases of trust, and regardless of intention it looks like the supervisor is using the CCTV footage to monitor the employee. The employee, were they made aware of this fact, would have to trust that the supervisor was in fact not monitoring and was instead pondering whether to use the footage as training material. The employee is well aware of how the action can be interpreted by others, and what is more important, they are aware that the supervisor is aware of how the action can be interpreted by others. The supervisor is willing to risk the trusting relation between them, regardless of whether their intention is pure, to analyse potential training footage. Moreover, the employee, when considering whether to trust the supervisor's testimony that they had good intentions, could consider the supervisor's disregard for their trusting relation as a part of the evidence they have regarding the supervisor's trustworthiness. Note here that if the supervisor would simply ask the employee for permission before watching the live CCTV (disclosed monitoring), then the case would fail to show what Keren wants it to show, namely, that there would be no relevant three-place trust to be undermined by monitoring.

The difference between the risk minimising effect of intentions one on hand and actions on the other, highlights the discrepancy of how the actions themselves (and not the intentions) are able

to mitigate risk in cases of trust, but the intentions still cause the undermining of trust. Moving forward with this line of thinking, consider the following example:

> HOLIDAY WEEKEND. Parents are going overseas to a remote cabin for a weekend and affirm their trust that the babysitter will take care of their child in the meantime. When they eventually arrive at the cabin, they come to two realisations. First, that they have doubts about their babysitter's trustworthiness. Second, that their phones have no signal, and there are no flights until Sunday afternoon.

This case involves intention to ignore preemptive reasons without the possibility of acting out the following precautionary measures. Imagine that if the parents arrived back and told the babysitter their intention, namely, that they wanted to take precautionary measures, but because of their situation they were unable to. The babysitter might not feel as if the parent's trust in her has been undermined, after all, they left their child with them for a weekend, which they would not do without trusting the babysitter.

The idea here is that for this to occur, something must have gone awry during the trust affirmation. If simply reflecting on the trust, without any new information, raises enough doubt to ignore preemptive reasons, then too much trust was initially affirmed. In that case, the trust should have been effectively undermined from the start. For example, if I tell my friend that I trust them to do something, and then renege on my statement a few moments later, it seems more likely that I was too hasty with my trust affirmation than my belief in my friend's trustworthiness had shifted ceteris paribus. The parents in HOLIDAY WEEKEND either trusted the babysitter with their child, or they did not, and a shift in intentions alone is not able to alter the degree of trust if the initial trust was built upon epistemically good foundations. This discussion is setting aside the fact that this is an ongoing definite trust case, so when the parents return home, they will engage in confirmation monitoring, which ignores preemptive reasons regardless.

To drive the point home, consider a patient going into operation where they will have to go under general anaesthesia. At first, they affirm their trust that the doctor will perform the

surgery without cutting into a tattoo that has great meaning to them, but during the operation the patient will not have any intention to respond to preemptive reasons.

Either we concede that one cannot trust anyone without being capable of intending to respond to preemptive reasons, or that one can trust without intending to respond to preemptive reasons. In the former scenario, one would have to admit that they stopped trusting their spouse the moment they went to sleep the previous night and started trusting them again when they awoke.[110] In the latter, Keren's claim that there is no trust without preemptive reasons fails.

Monitoring that Facilitates Trust

This rift between intention and action as they relate to precautionary measures has further complications. As we saw previously, precautionary actions cannot eliminate risk, only minimise it, and even then, it depends on the kind of action taken. To recap, even if one did not respond to preemptive reasons and instead decided to take precautionary measures, they would only affect the disutility risk of ¬Φ, and arguably not to an extent that would fully eliminate the risk. Even if one has the intention to ignore preemptive reasons, the intention does not by itself establish any risk-minimising effects. Contrast that with actions that can certainly minimise risk without the intention being the right kind of intention, i.e., the intention to neglect preemptive reasons. Such cases can be defined as cases of *pseudomonitoring*, in which the action and result of monitoring are realised de facto, without any intent to monitor.[111] Take for example a case in which a friend tells you, while you are playing snooker at a pool hall, that they recently won a snooker tournament. Unbeknownst to your friend, a recently installed plaque above the snooker table confirms what your friend is saying. Although you did trust your friend when they told you, and you harboured no suspicion or had any intention to monitor them, this plaque serves the same purpose as it would if you searched for evidence that corroborated your friend's testimony. Does simply noticing the plaque mean that you do not trust your friend any longer? On one hand, the role of trust in that particular instance is decreased, on the other, the two-place trust in your friend might increase because you have just confirmed that they told you the

[110] This relates to our initial discussion on definite and indefinite trusting.
[111] Note here that "unintentional" here only implies that there is no intention to monitor and does not facilitate the stronger claim that there can be no intention behind the actions that lead to risk minimising.

truth.[112] At the same time, it seems prima facie odd to claim that one can monitor without intention to do so, as when one is monitoring one is generally monitoring *for something*.

What is the relationship between intention and monitoring then? Emma Gordon (2022) finds that there are two prevalent theses that illustrate the connection between monitoring and trust. She calls the first one the strong tension thesis[113] and states that: "Trusting and monitoring are in constitutive tension with one another such that, when a trustor A monitors a trustee B with respect to task X (at time T), A *thereby* is not t rusting B with X at T" (Gordon 2022, 560). The second thesis, called the moderate tension thesis, concedes that there might not be a constitutive tension between trust and monitoring, but that monitoring is still oppressive, and erodes the trust at hand (Gordon 2022, 560). Formally stated, the moderate tension thesis states that: "Trusting and monitoring are in tension with one another in the sense that monitoring has an erosive effect on actual and possible trust relationships" (Gordon 2022, 560).[114]

Gordon initially challenges the moderate tension thesis by making use of two overdetermination[115] trust cases in which the trustor accidentally monitors the trustee, with the difference between them being that in one case the trustor is suspicious of the trustee while in the other the trustor harbours no such suspicion. Gordon's conclusion is that the trust erodes in the former example but not in the latter, i.e., in the latter case the trustor and trustee still have a good trusting relationship regardless of the monitoring that has occurred. This leads Gordon to reformulate the moderate tension thesis so that it better serves as a candidate for those who argue that monitoring and trust are always in tension, even if that tension is weaker than what is proposed in the strong tension thesis. The reformulation states that only the kind of

[112] It would be too strong to say that seeing the plaque would fully eliminate the need for trust, consider for example a modified case in which you have had a series of incidents of wrongly engraved plaques, in which case your friend's testimony would still be impactful. In short, whether evidence reduces the role of trust or completely eliminates it depends on the quality of the evidence.

[113] This view can be observed in writings by Nguyen (2022), Baier (1986), Jeremy Wanderer and Leo Townsend (2013), and Christiano Castelfranchi and Rino Falcone (2000).

[114] The moderate thesis view can be seen in some form in e.g., Hardin (1992) and Horsburgh (1960).

[115] The notion of overdetermination Gordon employs is adapted from the moral responsibility literature, where one can be morally praiseworthy for doing something even when the act was overdetermined (i.e., out of one's control). In this context it takes the form of monitoring that is practically unavoidable from the outset (Gordon 2022, 561). For clarity, I will use overdetermination in cases where there is intention to monitor but the monitoring happens accidentally, and pseudomonitoring when there is no such intention.

"monitoring *that manifests the trustor's intention to mitigate risk* has an erosive effect on actual and possible trust relationships" (Gordon 2022, 562).

Although Gordon ultimately offers a compelling reason for why even this reformulated moderate tension thesis should be rejected, it still serves as a good example of how the relationship between monitoring and trust is commonly understood. We can present an objection to the reformulated moderate tension thesis that is distinct from the one Gordon ultimately proposes. Recall that in the snooker case, seeing the plaque did not erode the trust you had in your friend's testimony, even if there was no need to trust their statement anymore. Gordon points out that because this kind of pseudomonitoring does not erode the trust one has, given that there is no risk-mitigation intention motivating the monitoring, then the revised moderate tension thesis does not hold.

I suggest that this is too hasty, and the conclusion, although convincing, is based on confusion. Namely, it fails to account for the distinction between two-place and three-place trust, which affects these cases of pseudomonitoring. In the snooker case, it is true that the pseudomonitoring of seeing the plaque does not erode your *two-place* trust in your friend, but what about three-place trust? An initial response is that there is no need to trust their statement anymore, and in fact it is not clear there is a way to keep trusting when one already knows.

When pseudomonitoring in a three-place trust instance provides overwhelming evidence of what was entrusted, the need for three-place trust is eliminated. In fact, it becomes impossible to maintain trust in that manner because the evidence is too strong. Consider that if you said to your friend after seeing the plaque "I trust that you are telling me the truth", when at the same time you know that they are telling you the truth, it would be disingenuous, as you are not revealing what you know. Claiming that you have a three-place trust in someone when you know enough to not having to rely on trust is simply one way to *act* like you are trusting. Imagine a typical case of trust acting, where someone claims to trust without trusting. Usually these kinds of cases depict the trust-actor acting like they trust the trustee while neither trusting nor knowing that the trustee *will* Φ, but we can still see that they are similar to cases in which one trusts the trustee to Φ, while knowing that the trustee has *already* executed Φ. Conversely,

it seems perfectly reasonable to proclaim to your friend that you trust them (in general) after seeing the plaque, without a hint of acting.

Gordon ultimately argues that the revised moderate tension thesis, that states that there is tension between trust and monitoring because monitoring performed with the intention to mitigate risk has an erosive effect on both actual and potential trust relations, is too strong (Gordon 2022, 562). In her view, the revised moderate tension thesis is too stringent because there are cases that show how monitoring, even when performed with the intention to mitigate risk, can not only fail to erode trust, but can even strengthen trusting relationships (Gordon 2022, 562). She proceeds to demonstrate how monitoring can be conducive to both two-place and three-place trust. I argue that when we factor in temporal elements of trust, we can perfectly explain what is happening in these cases, and why it looks like monitoring is beneficial to trusting relationships, while illustrating why this is not the case.

To demonstrate how monitoring can be beneficial to two-place trusting relations, Gordon begins by introducing an example of an indefinite two-place trust, in which a partner in a monogamous relationship accidentally discovers that their spouse has been unfaithful. Because the relationship was in many ways good, they agree to give the relationship a chance to recover, with the caveat that the affair stops and any contact with the affair partner is restricted to professional exchanges (as the affair partner was a coworker of the cheating spouse) (Gordon 2022, 562).

How can the trust in the relationship be rebuilt? Gordon references Shirley P. Glass (2007) who writes that openness, accountability, and honesty are the key factors in a successful restitution phase[116], and when a partner has been dishonest and deceptive, concrete evidence is needed to trust that the affair is truly over (Glass 2007, 325). One way to gather this evidence is by monitoring (Gordon 2022, 563). Gordon also points to a conclusion in a paper by Hertlein et al. (2017), that states that one way to reestablish trust is by demonstrating that the cheating partner has put an end to the affair and ceased contact with the affair partner. Testimony alone is not sufficient to demonstrate this fully, as testimony of this sort generally entails trust, but

[116] Snyder et al. (2007) define this period of time, in which a couple experiences fractured trust while trying to move past an affair, as *the restitution phase*.

the trust the spouse has towards the cheating partner has already been broken, which greatly reduces the evidential weight of the testimony (Gordon 2022, 563).

What kind of evidence is needed on this view to rebuild trust, and how is it obtained? One suggestion is to eliminate secrecy by subjecting the cheater to disclosed monitoring performed by their spouse, and having the cheater cooperate by being fully transparent and willing to help search for counterevidence, such as allowing full access to their phone, schedule, computer, etc. (Snyder et al., 2007, p. 304). If the cheating partner is unwilling to take on this level of accountability, Glass states that there is no reason to believe their word that the affair is over (Glass 2007, 325). After the cheating spouse has accepted that they will be monitored, the cheated spouse can take small gradual risks by slowly increasing the trust they put in their cheating partner (Gordon 2022, 563). It is important here to emphasise that on this view, "trust is built on the establishment of reliability and on the partner's ability to be predictable" (Hertlein et al., 2017, p. 329).

Gordon's conclusion about the role of monitoring in these kind of two-place trust cases is twofold. Firstly, monitoring can stop trust erosion and aid in rebuilding betrayed trust, and it is not merely a consequence of doubting but an active component of rebuilding trust. Secondly, monitoring in these cases must be executed in a controlled manner, with some specific time limit in mind, for it to be as effective as possible (Gordon 2022, 564).

Before moving on to three-place trust and beneficial monitoring, I would like to address these cases using the terminology introduced in this chapter. The aim is to show that although one can come to the reasonable conclusion that the moderate tension thesis should be rejected using Gordon's line of reasoning, there is a more intuitive way to think about the cases she presents, and account for them using the temporal elements of trust previously presented. For instance, we can interpret what is happening in the cheating case another way, namely, we can make a distinction between the ongoing two-place trusting relationship the couple had, and the trusting relationship that is created after the restitution phase. Between those two states we have the fractured trust that occurs after the affair was discovered, and the journey from that broken trust back to a state of ongoing indefinite trust.

A key difference here is that although Gordon claims that the trust is fractured, I suggest that it is broken in such an egregious manner that there is no trust remaining that can be an object of discussion. So on my view, the cheating example shows an indefinite ongoing two-place trust that is broken, and the couple decides to take steps to foster an environment in which a new trusting relation can be established. This new instance of indefinite trust must begin with a trust affirmation, followed by a series of therapeutic definite trust instances that are then able to therapeutically strengthen the trusting relationship.

This invites a discussion of what it means to monitor someone with whom you do not have a trusting relationship, and why therapeutic trust requires trust affirmation. In cases where there is no trust, one is free to gather evidence and reflect on that evidence along with any and all background information and experiences one has in their possession, until one finds themselves to be either trusting or not trusting. Without trust, monitoring simply describes normal epistemic practices of forming a belief. Glass claims that concrete evidence is needed to trust that the affair is truly over (Glass 2007, 325), and Gordon finds that one way to gather this evidence is by monitoring (Gordon 2022, 563). These claims hardly describe trust, they seem to be describing a state of knowing. See the following case:

> NEIGHBOUR. Alvey is the owner of a successful cybersecurity and risk management firm and is thinking about hiring a person to lead a new department. She knows that her neighbour, a former IT risk management consultant, is currently looking for work. Alvey thinks that the neighbour would be a good fit for the position because they are qualified, hard-working, and smart. Trust is also important due to the sensitive nature of the job, however, Alvey does not know whether the neighbour is trustworthy. She starts monitoring their neighbour to see if she spots any signs of dishonesty or deception. This includes asking the neighbour questions she already knows the answer to, to see if the neighbour will lie. She reflects on previous encounters and tries to find out if they have a drinking problem, whether they like gambling, and so forth. One day she even leaves her wallet on her lawn, to see if her neighbour will return it to her. After monitoring their neighbour for a few weeks, Alvey is convinced that her neighbour is trustworthy, she might even claim that she knows that their neighbour is an upstanding person, with nothing to indicate that they are untrustworthy.

Even in an extreme case like NEIGHBOUR, it is difficult to claim with certainty that the neighbour would be upset about being monitored before being offered the job. After all, there was no trust between Alvey and the neighbour, and regardless of the ethical considerations of setting up these kinds of *trust-traps*, the neighbour is in no position to feel betrayed by having been observed.[117] The reason is that there was no trust at the time of the monitoring,[118] and these kind of pre-trust precautionary measures are performed to get a better grasp on whether or not one should affirm their trust in someone.

Now that we have gotten an idea of what it means to monitor someone without a trusting relationship to go with it, we can focus on the notion that trust affirmation is needed to therapeutically trust someone. Keren (2014) provides a convincing argument for why the cases that are usually lauded as paradigmatic cases of therapeutic trust are in fact not entirely as they seem. Specifically, Keren argues that these therapeutic trust cases, such as a teenager taking care of the house while the parents go away for the weekend, or a neighbour watering one's plants, do not possess one of the essential characteristics of therapeutic trust as it has been portrayed in the literature at large, which is the phenomenon of trusting someone without believing that they are trustworthy. To demonstrate this, Keren introduces two cases of what he considers to be *real* cases of therapeutic trust that look wildly different from the cases we have come to know as the paradigmatic therapeutic trust cases. One of those cases involves trusting someone that is perceived to be untrustworthy from reasons provided by a third party,[119] and the other presents a trustee that is trustworthy, and because of that they can complain if they are not really trusted.

The latter case is of interest to us here, which can be described follows. Mr. Barnes is a faithful husband with a suspicious wife that has always doubted her husband's faithfulness. They agree to go to counselling to save the relationship, as her doubts had almost led to their marriage collapsing (Keren 2014, 30). After the counselling, Mrs. Barnes has taken on an affective attitude of optimism and manages to act like she trusts her husband (Keren 2014, 30). Even so,

[117] Although they might find the behaviour creepy and disconcerting.
[118] No relevant trust at least. One could argue that we trust each other to follow societal norms, but without trust affirmations these can be considered cases of reliance in lieu of trust.
[119] In Keren's case this third party has an authoritative status. For further reading see Raz (1990).

Mrs. Barnes still has serious doubts about her husband's goodwill and faithfulness and believes that her doubts are still justified. If Mr. Barnes would come to know this, even if she did not express her doubts, it looks like Mr. Barnes could complain that his wife does not really trust him. What is more, Mrs. Barnes could not really respond by saying that she in fact really trusts him, but that she also has serious doubts about his trustworthiness (Keren 2014, 30).

Contrast this case with typical therapeutic trust cases. The difference between Keren's therapeutic trust cases, and general therapeutic trust cases we are familiar with (like the teenager watching over the house), is that in Keren's cases, there is absolutely no belief of trustworthiness, while in the general cases, we find that there is *some* limited belief that the trustee is trustworthy at least to a degree. I am inclined to agree with Keren. It seems to me that for therapeutic trust cases as they are commonly understood, such as the house-sitting teenager case, some limited belief, that involves a degree of trust at the onset of trust, is required for therapeutic trust cases to have the aim of building trust.

Note the therapeutic trust relationship that emerges in Gordon's cheating partner case, although partially effective at establishing a new two-place trust between the cheater and the cheated spouse, is itself a three-place trust. Recall that when the couple made the decision to mend their trusting relationship, using monitoring and therapeutic trust, restrictions were placed on the cheating partner. Specifically, that they end the affair and limit contact with the affair partner to purely professional exchanges (Gordon 2022, 563). These restrictions serve as two separate ongoing cases of three-place trust, namely (1) the cheated spouse trusts cheater to end the affair (definite trust), and (2) the cheated spouse trusts the cheater to limit contact with the affair partner to purely professional exchanges (indefinite trust). Gordon's case is convincing precisely because it employs various features of both definite and indefinite instances of three-place trust, while at the same time presenting the case as being a case of two-place trust. One way this is utilised is by claiming that the role of monitoring in the restitution phase should only be performed with a kind of time limit in mind to be maximally effective (Gordon 2022, 564). Note here that this kind of a time limit suggests that monitoring and trust are in fact at odds with one another, and as the trust increases, monitoring must be decreased if the increased trust is supposed to hold.

Even if we would concede that monitoring (2) would be acceptable with a time limit in mind, the same cannot be said for (1). There is no time limit for monitoring that "the cheater ends the affair"; it is a case of an ongoing definite trust that is resolved only if the cheater in fact ends the affair and confirmed only in case the cheated spouse engages in confirmation monitoring to gather evidence that the cheater has done as entrusted. To further support this claim, consider a scenario where the cheated spouse would monitor their partner with regards to (1), and at no point would they gather any evidence that their partner had in fact ended the affair (in the form of a break-up text, or seeing that they are no longer connected on any social media platform). At some point the time limit would be up, and the cheated spouse would no longer be able to monitor their partner without damaging the trust they have. It looks like the monitoring (1) did nothing to mend the trust between them. An immediate objection that comes to mind is that this is just a matter of phrasing. It is easy to change (1) into something along the lines of (1)* they do not continue the affair[120], but the point still stands, these would be two instances of an ongoing indefinite three-place trust.

The revised moderate tension thesis can still challenge to the claim that monitoring can be beneficial when it is limited to three-place trust cases (Gordon 2022, 566). The revised three-place moderate tension thesis states that monitoring and trust are in tension because monitoring, when performed by the trustor with the intention to mitigate risk, has an erosive effect on three-place trust relationships (Gordon 2022, 566). Even with these further restrictions to the tension thesis, Gordon finds cases that seem to show that monitoring can sometimes be conducive to three-place trust, thus opposing the claim that there is necessarily tension between trusting and monitoring. The way she goes about this is by arguing that a general two-place monitoring in the original cheating case can promote three-place trusting. The example she uses resembles the previous case, where the cheater agrees to disclosed monitoring to provide the cheated spouse with evidence about their trustworthiness. The difference between this case and the two-

[120] Given that an affair that is on and off constitutes a singular affair, and not multiple affairs. If that is indeed the case, then we must resort to a less elegant statement such as "they will not behave in any way that constitutes an affair", where affair can be defined as inappropriate physical, emotional, or mental relationship with someone besides one's partner. Note that the term "inappropriate" does heavy lifting here and can cover edge cases such as polyamorous relationships with a specific ruleset, in which case inappropriate behaviour just means behaviour that goes against the spirit of the rules that have been placed.

place trust case, is that in this case the cheated spouse trusts cheater to perform some specific action, such as coming straight home after work (Gordon 2022, 567).

The claim here is that if the cheated spouse has engaged in disclosed monitoring for some time, then they are more likely to trust the cheater in these three-place trust cases ceteris paribus than they would without the disclosed monitoring. Monitoring is thus supposed to provide evidence of a person's trustworthiness that can have "an impact on the likelihood of three-place trust in similar future cases" (Gordon 2022, 567). Disclosed monitoring in particular can thus facilitate three-place trust by fostering an environment that makes it more likely, even if further monitoring might undermine trust once the trust has been placed.

Now, from our discussion on the two-place trust examples we see that this view runs into similar problems as before with one slight variation. Again, the issue is that there is no trust involved at the start of the restitution phase that follows the discovery of the betrayal. We see, in much the same way as before, that the kind of monitoring that happens following the counselling is just standard evidence gathering as one would find in any generic doxastic scenario. The previously established trusting relationship does not impact this new relationship except for the expected cooperation of the cheater. This again leaves us with Gordon's view only showing that monitoring is not in tension with trust when the people involved do not have any trust between them.

The way to oppose this conclusion is to say that there is some degree of trust left in the relationship that suffices to make the claim that monitoring is not in tension with trusting. As a first response, it seems unlikely that there is any meaningful trust left. Even if the cheated spouse can still claim that they trust their partner in some ways, having shared many years together and developed a strong bond, it seems naïve to suggest that she trusts her partner to be faithful in the immediate aftermath of discovering that her partner has been unfaithful. Furthermore, such a discovery would, in most cases, lead to serious doubts about other aspects of their relationship. For one, the cheated spouse might realise that their partner is a much better liar than they previously assumed, and the amount of deception needed to maintain an affair would further fuel those doubts. It is not fantastical to assume that this level of deception would severely affect multiple aspects of their relationship, many of which have nothing to do directly

with the affair. To illustrate, a cheated spouse might not even trust their partner with their children and might say something like: “I don’t even know who you really are”, even if the cheating partner would have consistently demonstrated great parenting abilities throughout their relationship.

Although we agree with the claim that not all monitoring is in tension with trusting, as we have seen with confirmation monitoring in definite trust cases, it is not clear that this is the case in indefinite trust cases such as “I trust that my partner will never cheat again” unless we can claim that indefinite trust is only a series of definite trust cases. If that would be the case, then we could argue that:

(P1) Confirmation monitoring is a kind of monitoring.

(P2) Confirmation monitoring is not in tension with trust.

(C1) Not all monitoring is in tension with trust.

(P3) Definite trust cases involve confirmation monitoring.

(P4) Indefinite trust cases are a series of definite trust cases.

(C2) Indefinite trust cases involve confirmation monitoring.

Gordon claims that specific three-place ongoing definite trust situations can improve and indeed create a general two-place trusting relation, which could be used to defend (P3). However, this would mean that she would have to accept that monitoring for one thing leads to trust unrelated to what was being monitored, which brings us back to the notion of pre-trust affirmation monitoring, i.e., that pre-trust monitoring is just standard evidence gathering without any trust considerations at hand. Can S be trusted to Φ, while being monitoring for Φ+, where Φ+ can be thought of as the possible actions one could generally be trusted with in a two-place trust relation, including the specific Φ that S was trusted with in the three-place trust?

If we allow specific instances of three-place ongoing definite trust to generate two-place trust, then the confirmation monitoring would not serve to increase the relevant three-place ongoing definite trust, rather, it would be an opportunity to find evidence that would facilitate a general two-place trust, such as whether the trustee is generally an honest, reliable person. To clarify, consider the following example:

> OFFICE TRUST. Amy only wants to have friends she can generally trust. A new coworker just started at Amy's job, and they get along quite well. Amy senses that the coworker desires to make friends, as they recently moved to the city. Before going out of her way to befriend the coworker, Amy decides to monitor the tasks she trusts her coworker with. The tasks are limited to their work, and the trust is of a limited degree that only relates to the workplace.

After monitoring the coworker for some time, it looks like Amy has successfully gathered evidence about whether the coworker can be trusted to do as entrusted at work, such as taking on small tasks for Amy, and not betraying their professional relationship by, e.g., stealing her ideas and presenting them to the board as their own. In contrast, it is not as clear that Amy has more reason to believe that her coworker is generally trustworthy, and can thus be trusted with secrets, her ailing dog, and to be there for her in times of need, than without the monitoring.

If Amy cannot establish a two-place trusting relation with her coworker, it raises the question whether and how two-place trust can come about. After all, it looks like one can only generate or increase three-place trusting relations that are the same or similar to successful previous instances of three-place trust. However, note that that the argument was initially presented as a way to defend the claim that trust and monitoring can coexist without tension in some indefinite trust cases. This does not have any bearing on the way definite trust is therapeutically increased using confirmation monitoring, or how one maintains indefinite trusting relations.

To summarise, the main argument here has been that although we agree with Gordon's general conclusion, that monitoring can be beneficial to trust and they are thus not in always in tension with one another, we disagree with the reasoning. Namely, confirmation monitoring is essential to therapeutic trust cases, but detrimental in regular indefinite trust cases, even when it cannot

minimise the impact of betrayal. The friction between the view presented here and Gordon's view stems from a different understanding of the restitution phase of the affair example. On my view, there is no trust and there can thus be no monitoring as it is commonly understood, rather, the behaviour the cheated spouse is engaging in is more akin to evidence gathering. On Gordon's view however, the trust is not broken but fractured, and monitoring is a way to mend the trust.

As a final note, Gordon presents Wanderer and Townsend's (2013) nanny-cam case, in which parents leave their child with their nanny but cannot help themselves and decide to monitor the nanny through their security system, as an example of a three-place trust that seems "obviously wrecked" by the parent's monitoring (Gordon 2022, 566). We have demonstrated how this kind of three-place ongoing definite trust needs to be confirmed by confirmation monitoring for it to have therapeutic benefits. In the nanny-cam case we see that the parents will eventually return home and obtain some evidence about whether the nanny did as they were entrusted to do. Note here that it is unlikely that the parents will be able to fully confirm their trust in this case. That is, although they can check their child for obvious signs of neglect or abuse, it is difficult to imagine that they would be in a position to monitor them to the extent that they can be fully confident that the nanny did everything as entrusted. For example, it is exceedingly difficult for the parents to know whether the nanny yelled at their child at some point.

One consequence of this partial sort of monitoring is that if the nanny-cam case was therapeutic, i.e., the parents do not trust the nanny but foster an environment to give the nanny an opportunity to showcase that they are trustworthy, thus establishing a way to therapeutically trust the nanny,[121] then this sort of partial monitoring would only be partially successful at therapeutically increasing the trust. To see why, consider that they have a three-place (therapeutic) trust in the nanny to take care of their child, in which "taking care of" entails various responsibilities, such as not raising their voice, providing healthy food, be willing to entertain them, and so on. However, the parents cannot use confirmation monitoring to gather

[121] In addition, the parents hope they find that their trust in the nanny increases incrementally by repeated interactions, that trust increase might result in the parents being comfortable leaving their child with the nanny for longer periods of time or going further away from their home.

evidence about these things, being only positioned to evaluate some parts of what the nanny was entrusted to do. For example, maintaining a clean house, feeding the child (depending on how long the parents were gone), not abducting the child, and not engaging in abuse that leaves noticeable signs.

This kind of evidence from confirmation monitoring surely counts, but it does not eliminate the risk that the nanny will not yell at the child. Conversely, monitoring the nanny through the security system would alleviate those concerns, but that would not be trusting. It is too strong to say that monitoring and trust are in tension in the nanny-cam example, but it is plausible to claim that some monitoring is in tension with trust in the nanny-cam case (and other definite three-place trust cases where the confirmation monitoring is only partial). Recall that in some definite three-place trust cases the confirmation monitoring can fully confirm the resolved therapeutic trust, in which case the therapeutic trust is maximally effective.

The Tension Between Rationality and Trust

The temporal elements of trust that have been employed throughout this chapter can also explain why some examples of trust seem to involve trust as a rationally appropriate attitude, while others do not. Wanderer and Townsend (2013) ask whether it is rational to trust. In their paper, they introduce the nanny-cam case and point out that it looks like there is tension between trust and rationality (Wanderer and Townsend 2013, 1). They find that there are three claims that, when conjoined, form the tension between trust and rationality.

First, that it is central to the idea of trust that it is a cognitive attitude in the sense that it must involve something akin to a belief about a person's trustworthiness, or at least that they will not betray us (Wanderer and Townsend 2013, 1). Second, that the primary set of norms that govern whether a cognitive attitude is rational is associated with evidentialism (Wanderer and Townsend 2013, 1). Finally, that trust is in tension with evidential norms because trust is "formed and maintained in ways that extend beyond that which is supported by evidence" and moreover because the forming of and maintaining an attitude of trust is itself resistant to the weighing of evidence regardless of what the evidence supports (Wanderer and Townsend 2013, 2).

These three claims lead to the conclusion that trust is not a rationally appropriate attitude (Wanderer and Townsend 2013, 2). Wanderer and Townsend explore various possibilities to avoid this conclusion, e.g., rejecting any one of the claims, but find that none of the rejections are convincing. Furthermore, conceding and simply allowing the tension by endorsing that trust is not a rationally appropriate attitude would lead to the undesirable result that it is possible to acknowledge that some attitude is irrational while at the same time claiming that it is the right thing to do (Wanderer and Townsend 2013, 11). In a final attempt, they accept the tension, but on their own terms. Instead of assuming that the tension between trust and rationality needs solving, they reflect on the phenomenology of trust and find that perhaps the tension is genuine and ineliminable from a practical standpoint (Wanderer and Townsend 2013, 11). Additionally, they argue that rational doubts can make trust "all the more commendable", as it means that someone still trusts even when harbouring serious doubts, although they do not claim that all cases of good trusting need to be conflicted (Wanderer and Townsend 2013, 12).

Using temporal elements of trust, we can argue that there is less tension between rationality and trust in definite trust cases than in indefinite trust cases. Then we can demonstrate how Wanderer and Townsend's final attempt is not convincing, even for indefinite trust cases, by considering the PHONE COUPLE case already introduced, along with a new line of reasoning that emphasises further why doubts, as opposed to monitoring, are in fact detrimental to trusting relationships.

I have argued that for definite trust cases to be deemed as completely successful cases of trust they must involve a sort of confirmation monitoring, either unintentional or deliberate, disclosed or undisclosed, if they are to increase the trust and not simply maintain it. It follows that cases of definite trust are in less tension with rationality, as they not only allow but also require a specific set of evidentialist norms, than cases of indefinite trust that do not seem to withstand monitoring to any degree. Furthermore, it is crucial to clarify whether precautionary actions are being taken before or after a trusting relation has begun. In cases where trust is imminent but not established, monitoring and other such rational precautionary measures are welcome and do not put a strain on the trusting relation that can then be doxastically formed.

Definite trust cases are not necessarily in tension with rationality when coupled with confirmation monitoring, but what of indefinite cases of trust? If we examine Wanderer and Townsend's (2013) attempt to solve the tension by accepting it as genuine and ineliminable only as an attempt to solve the tension in *indefinite trust cases*, wherein evidence gathering such as monitoring can certainly undermine trust, it still comes up short. They conclude that the tension between trust and rationality can be commendable, citing the phenomenological and practical experience of trust wherein vexed trust is harder to maintain than unvexed trust. However, we can recall the PHONE COUPLE case, in which two individuals maintain indefinite three-place trust in the other, but one experiences vexed trust while the other does not. In that case we concluded that unvexed trust is in fact stronger than vexed trust because reflecting on trust ceteris paribus signals that the trust has been threatened (Nguyen, Trust as an unquestioning attitude 2022, 15).

Vexed trust indicates doubts about the trustee's trustworthiness. Whether those doubts are substantial enough to lead the trustor to stop trusting does not have a bearing on the level of trust involved. Conflicting intuitions about a particular trust relation do not make that trust more commendable because although it shows that, if the doubts are serious, the level of trust must be high enough to prevail over the doubts, it does not show that the absence of such doubts makes the trust any less commendable. In fact, it is easy to argue that doubts make trust less commendable than trust without doubts. To see why, consider again PHONE COUPLE and notice that if something were to happen that increased the doubts that they both had about each other's trustworthiness, such as an unlabelled bouquet of roses from an anonymous sender, then the vexed trust individual would be more likely to stop trusting their partner and might engage in further monitoring behaviour such as going through their phone without their knowledge. The unvexed individual, after reflecting on this new evidence, might understandably lower their trust in their partner, after all they have engaged in undermining behaviour by taking the bouquet of roses into account when reflecting on their trust. However, without any other doubts in their mind about their partner's trustworthiness, these doubts are not as likely to have a deciding effect on whether they trust their partner as they would if they had harboured serious doubts about their partner's trustworthiness to begin with.

Conclusion

In the first section we examined trust within a three-place trust framework, where A trusts B to perform a certain action, Φ, and A relies on B to fulfil this trust. A distinction was made between trust affirmation and ongoing trust. Trust affirmation can be characterised by A's initial belief towards B's trustworthiness, doxastically formed using evidential norms. Before the trust affirmation takes place, precautionary actions of trust can be freely executed without undermining the trust that follows. However, when trust is being established during the trust affirmation itself there is no time to engage in precautionary measures such as monitoring, searching for evidence, or rational reflection.

Trust affirmations are followed by ongoing trust, where the trusting relation continues over a period of time. An example of this is a couple who trusts each other to be faithful in their relationship. We discovered that not all cases of ongoing trust react to precautionary action in the same manner. To clarify why that is, definite and indefinite notions of trust were introduced.

Definite trust involves cases in which trust has a clear terminus, such as trusting someone to perform a specific task within a set timeframe. We found that therapeutic trust cases must necessarily be instances of definite trust, and the concept of confirmation monitoring was introduced. Confirmation monitoring is monitoring that occurs at the end of definite trust instances. Therapeutic trust cases can be successful as long as they aim to build trust, and we argued that confirmation monitoring is necessary for any therapeutic increases in trust. As a result we further argued that successful therapeutic trust cases are always cases of definite trust.

We explored the connection between precautionary actions and temporal elements of trust and how they are tied to risk. A distinction was made between the risk of being betrayed and the risk of disutility in cases where the entrusted action was not executed. We analysed the risk-mitigating effects of precautionary actions and whether temporal elements impacted those actions in any way. We argued that precautionary measures can partially guard against the risk of the entrusted action not being completed, but that it cannot guard against the risk of being betrayed. Furthermore, we concluded that to fully guard against risk, one must partake in precautionary actions that involve intervening to an extent that they eliminate the need for trust

altogether. This reflects a complete lack of trust, rather than doubts, and is incompatible with trust in a way that milder forms of precautionary actions are not.

We then explored the implications of the distinctions introduced in this chapter on an existing trust account, namely, Keren's doxastic trust account with preemptive reasons. Keren claims that reasons for trust are second-order preemptive reasons against engaging in precautionary actions like monitoring or reflection. However, when considering how preemptive reasons and precautionary measures affect different kinds of trust instances in different ways, we found that preemptive reasons are not always present, and when they are they are not necessarily suited to minimise risk, thus having minimal effect on the trusting relation.

We concluded that, although preemptive reasons, as they are commonly understood, are not present during trust affirmations, they are essential in ongoing indefinite trust cases, with some caveats. We further argued that preemptive reasons must necessarily be ignored if ongoing definite trust cases are to be successful. We considered objections that explored the relation between risk and precautionary actions, and how trust and precautionary actions relate to intention and action.

In the following section, we examined Emma Gordon's view that monitoring can facilitate trust. Although we agreed with Gordon's general assessment that monitoring can be beneficial to trust and, therefore, that there is less tension between trust and monitoring than commonly thought, we came to the same conclusion for different reasons. I argued that when trust is rebuilt after being broken, it must be reaffirmed, thus causing a new trusting relation to begin. One consequence of this is that it explains why monitoring seems compatible with trust in the way Gordon claims, namely, because in between the terminus of the previous ongoing trust and the trust reaffirmation is a period without trust. During this period, epistemic agents are free to monitor because there is no trust to undermine. I still argued that the overarching claim is correct, that monitoring can facilitate trust, because confirmation monitoring is a necessary feature of successful therapeutic trust increases and definite trust cases more generally.

Finally, we argued against Wanderer and Townsend's attempt to reconcile trust and rationality. They accepted the tension between trust and rationality by claiming that the tension is genuine

and ineliminable, and that rational doubts can make trust all the more commendable. In our response, we showed that the tension in definite trust cases is limited due to confirmation monitoring. This monitoring, whether accidental or deliberate, plays an important role in the success of definite trust cases. In contrast, indefinite trust is undermined by monitoring in the way Wanderer and Townsend suggest. We then argued that vexed trust is less desirable than unvexed trust. Vexed trust is a result of doubts that one harbours, and reflecting on trust does not strengthen it; on the contrary, it undermines it. Additionally, we showed that in cases where two individuals trust, one vexed and the other unvexed, less is needed for the vexed trustor to stop trusting, further indicating that unvexed trust is less fragile.

Chapter 4: Conditions for Justified Epistemic Gatekeeping

Abstract. There is a distinct difference between acquiring knowledge and distributing it. Greco (2020), in his book *The Transmission of Knowledge,* points out that having different social norms for knowledge acquisition on one hand, and distribution on the other. This allows epistemic group agents to have cheap knowledge transmission between themselves using distribution norms, while the assessment of outside testimony, a sort of gatekeeping, requires stricter acquisition norms. This chapter presents two problems for gatekeeping. First, the negative consequences of believing falsehoods are exacerbated through gatekeeping (POISONED WELL). Second, the likelihood of the group being deprived of knowledge is elevated (KNOWLEDGE DEPRIVATION). One way to defend against these problems is to assign more resources to the gatekeeping process, i.e., increasing the emphasis on acquisition norms at the cost of distribution norms. I show that any reduction is epistemic labour increases the risk of the gatekeeping problems. I further argue that gatekeeping increases belief uniformity within groups and show how that relates to the problems. Do we have to abandon gatekeeping or is there a way to preserve cheap knowledge transmissions? I propose three conditions of justified gatekeeping that, if met, can lower epistemic labour costs while minimising the epistemic harm of gatekeeping failures. Finally, we show how groups within expert communities are in a good position to implement justified gatekeeping.

Motivating Gatekeeping

Gatekeeping is, at its core, the process of controlling information as it moves through a gate (Barzilai-Nahon 2008, 1496). It has been applied in various ways in epistemology[122], most of which only share the bare essential feature of gatekeeping, i.e., its function to keep some things out while letting other things pass through, based on some specified criteria[123].

[122] For some recent examples in epistemology, see Greco (2020), Ulatowski (2022), Kwong (2023), Elgin (2020), and Henderson (2011).

[123] If we disregard the various considerations that dictate how, and why, a gatekeeping process is employed then it looks like gatekeeping as a mechanism is exempt from value judgements and can prove to be epistemically and/or morally good or bad depending on application.

The purpose of this section is to shed some light on the nature of epistemic gatekeeping as it relates to groups[124], the potential benefits of having some kind of gatekeeping in place, whether that be with or without dedicated gatekeepers, and the problems that can arise when gatekeeping has been employed. Finally, after examining scientific communities in particular, some conditions on gatekeeping as it relates to groups will be laid out to minimise the impact of the gatekeeping problems introduced in this chapter.

The core function of gatekeeping is relevant to vastly different disciplines, and it manifests in different ways, which can make it challenging to determine exactly what the term gatekeeping is meant to invoke. Gatekeeping as a term was initially coined to describe how some group members can exert control over what enters the group and what gets left out (e.g., what food is purchased for the household) by having the authority to make decisions on the group's behalf (Lewin 1943, 37-40). It has since been used to describe the screening procedures used in academic programs, how news has been curated by editors, and the selective exposure of social media.[125]

To further complicate matters, the concept of gatekeeping, as it is commonly used in everyday language, refers to the act of excluding individuals from an interest by restricting or discourage other's participation, enjoyment of, or identification with something, particularly an activity or interest which the gatekeeper shares (Oxford English Dictionary n.d.). More generally, it involves situations where a group or an individual is in a position of power to determine whether someone else should have access or rights to a particular identity or a community (Friedman 2023).

The concept of gatekeeping used throughout this chapter will refer to epistemic gatekeeping as it is exercised in John Greco's recent works (2015, 284, 2020, 39) to explain how the standards

[124] There are various ways in which philosophers have referred to groups as agential entities in the literature, such as corporate agents (List and Pettit 2011), collective agents (Searle 1990), group agents (Tollefsen 2015), and plural agents (Helm 2008). The differences between the accounts listed here do not impact the structure or arguments made in this chapter so for simplicity's sake groups that fit the definition proposed here will be simply referred to here as "groups".

[125] See e.g., Swank and Smith-Adcock (2014), White (1950), and Welbers and Opgenhaffen (2018), respectively.

of knowledge transmission are lower when transmissions occur within groups.[126] In short, Greco (2020) proposes that different social situations necessitate different epistemic norms, some of which can serve the role of gatekeeping in groups, or "a community of information sharers" (40).

To illustrate, imagine a group of close friends that are on holiday visiting a foreign city they have never been to before and they need directions to get to their hotel. One member of the group approaches a passerby and asks for directions. In doing so, they employ norms of knowledge acquisition which serve a gatekeeping function; they decide to ask someone who appears to be familiar with the city, is friendly, and so on. Furthermore, if the directions the group member received were strange or unbelievable, they would find someone else to ask. This process will generally match our understanding of what good epistemic practices entail. When the group member has received testimony that meets the standards of their knowledge acquisition norms and consequently forms a true belief about the location of the hotel, they share the information with the rest of the group. It then becomes clear how the norms of knowledge transmission have shifted; the friends trust one another and have ample reason to think that the knowledge acquiring friend would not share what they learned unless they believed it to be true. The knowledge is thus easily shared between the members of group, as the hard epistemic work of acquiring knowledge has already been done by the initial knowledge acquirer, allowing the rest of the group to get the knowledge for cheap.

Greco does not claim that there must be designated gatekeepers, so for example, the friend that asked for the directions to the hotel does not have a specified role as a gatekeeper; any member of the friend group could have asked for directions. Instead, gatekeeping depends on the kinds of knowledge norms that are used in testimonial exchanges which differ depending on whether the testimony comes from a speaker within the group or outside the group, where the hearer is a group member. So for example, one kind of gatekeeping could have a designated gatekeeper, which is the group's only contact with the rest of the world, who would carefully examine all incoming information using acquisition norms before transmitting it to the rest of the group.

[126] This notion can be traced back to Edward Craig (1990), in his book *Knowledge and the State of Nature* where he states that the purpose of knowledge is to flag good sources of information to serve the informational needs of the group.

Another kind of gatekeeping, which is more practical, is that all members rely on acquisition norms when receiving information from outside the group, while relying on distribution norms among themselves.

Another way to think about gatekeeping is to describe it as the role of acquisition norms in groups *when* there are distribution norms being used. To clarify, imagine a group that would rely on acquisition norms to the same degree regardless of whether they were having a testimonial exchange intra-group or outside of it. In that case it could be said that the group is gatekeeping knowledge, but that knowledge is just as gatekept within the group, as there is no cheap transmission available within the group after the initial knowledge acquisition. Gatekeeping, as a concept to describe the mechanism of allowing only high-quality information to get inside a group which can then be easily shared between its members, thus requires not only the acquisition norms of knowledge, but distribution norms that are appropriate because the knowledge has already been acquired using acquisition norms.

What does this kind of norm-driven gatekeeping look like? When an activity is performed with the aim of acquiring knowledge, e.g., gathering information, that is to be introduced "into the community of knowers in the first place", the norms that govern that activity effectively play a gatekeeping function (Greco 2020, 39). When the activity is performed with the aim of distribution knowledge within the group then the governing norms are meant to facilitate easy knowledge transmission. For clarification, these norms can be summarised as follows:

> **Norms of group knowledge acquisition for groups:** serve a gatekeeping function, "they exert quality control so as to only admit high-quality information in the social system" (Greco 2020, 39).
>
> **Norms of knowledge distribution for groups:** serve a distribution function, "they allow high quality information already in the system to be distributed as needed throughout the system" (Greco 2020, 39-40).

These two kinds of norms turn out to be different from one another because it should be more difficult for information to enter a system than to be shared within it. Thus, information must

withstand a quality control process before it enters a system, which makes subsequent transmissions within the group possible without exercising the same kind of resource-heavy quality control (Greco 2020, 40).

Greco also states that testimonial knowledge itself has two different kinds, i.e., testimony can either serve as the acquisition function of the concept of knowledge, or as the distribution function of it (Greco 2020, 41). For example, the testimony of the passerby is a part of the knowledge acquisition process, while the group member's testimony to their friends is a part of the knowledge distribution process. Both acts have their respective role in making sure epistemic groups are spending the appropriate resources to transmit knowledge. So, the norms of testimony here can roughly be categorised as knowledge acquisition norms, which serve the role of gatekeeping, and knowledge distribution norms, which allow for easy knowledge distribution.

To summarise, Greco proposes that groups can sometimes rely on knowledge distribution norms instead of acquisition norms when communicating with other group members, which lowers the barriers for knowledge transmission. To clarify, there is less demand to be sceptical when evaluating the testimony of the members of your group, e.g., the friends that were on holiday do not need to engage in the same rigorous epistemic process as if they were listening to the testimony of a stranger. Although both norms are relevant to some degree in most testimonial exchanges, it is apparent that some groups allow for the use of knowledge distribution norms in lieu of acquisition norms to a greater extent than would be possible without having an established group. Gatekeeping can thus encourage information flow intra-group while maintaining the epistemic quality of the information that is being introduced to the group.

Belief Uniformity and Group Agency

I argue that gatekeeping can also lead to increased *belief uniformity* among group members, which can be defined as follows:

> **Belief uniformity:** the extent to which the total body of knowledge held by a group of individuals overlaps.

When the information flow to a group is curated through a gatekeeping process that operates under some prearranged restrictions the body of total knowledge being employed by group members becomes increasingly homogenised.

It has been argued that "[…] within epistemic collaborations, joint commitments also play a role too crucial to ignore: To interact in a coordinated manner, group members need to both agree on and hold fast to certain assumptions or even a specific body of knowledge" (Palermos 2022, 54). Additionally, decisions can be more difficult to make in groups than at an individual level, as group members might not agree with each other doxastically (Eder 2020, 185). Even though a diverse body of beliefs in a group can make it harder to reach a conciliatory conclusion with regards to any decisive action, it is not impossible. Epistemic compromises can be reached by various means, including aggregation methods of epistemic states, democratic processes, and pragmatic concessions by group members (Eder 2020, 185). What these procedures have in common is that they aim towards reconciling the total body of beliefs within the group to a degree where the group members are homogeneous enough in their beliefs that the group can rightfully be said to hold the relevant beliefs according to the general standards in the literature of what is required of group members for the group to hold a belief.

This can further be seen when we look at group agency and intentions. A common view[127] in philosophy is that agency involves, at the very least, a system acting in some way in its environment to achieve a set goal (Lewis-Martin 2022, 5). One line of argument made by Tollefsen (2002) for groups being intentional agents, that focuses specifically on the goals and norms of groups, is that it can explain the actions of organizations as they relate to their beliefs, intentions, and desires being successful (397). She claims that groups must form a coherent whole, where the performance of joint actions through "[…] group ends, shared intentions, joint commitments, or we-intentions" in and of itself might be the manner in which groups form and

[127] Lewis-Martin (2022) lists Barandiaran et al. (2009), List and Pettit (2011), Tuomela (2013), and Tollefsen (2002), as having accounts of group agency that loosely follow this definition. Some of those accounts, like the one List & Pettit present, add some additional criteria while others, like Tollefsen's, emphasise group goals.

maintain their agency as persistent entities that are unified in one way or another (Tollefsen 2015, 47, Lewis-Martin 2022, 5).

Having a more unified total body of beliefs within a group should theoretically aid in maintaining a group's agency, as the group members will be more likely to share intentions that encourage joint actions towards some goal which solidify the group's agency. Consider that a group composed of individuals that each have their own vastly different set of beliefs would in some instances face serious challenges trying to perform joint actions, as their beliefs would influence their intentions.

The idea of belief uniformity is complementary to Jennifer Lackey's (2021) conditions for justified group[128] belief found in *The Epistemology of Groups*; a group justifiedly believes that *p* iff:

1) a significant percentage of the operative members of it justifiedly believe that *p* and if the bases of those believes were added together they would yield a belief set that is coherent, and
2) full disclosure of the relevant evidence regarding the proposition that *p* along with rational deliberation (in accordance with both individual and group epistemic normative requirements) within the group about the evidence would not result in further evidence that would yield a total belief set that would not make it sufficiently probable that *p* (Lackey, The Epistemology of Groups 2021, 97).[129]

To be clear, this is not to say that increased belief uniformity necessarily increases justification for beliefs, rather, it indicates that a group with high belief uniformity will have a more coherent total set of beliefs, i.e., a larger share of the group's beliefs will be justified than otherwise.

[128] Lackey presents a general feature of groups that can be used to identify and capture the kinds of groups she is ultimately interested in, namely, all groups that are subject to normative evaluation (i.e., both epistemic and moral normative assessments) (Lackey 2021, 11-12). This arrangement can be summarised as follows: If a group can be responsible for something, then that group should be considered relevant to her discussion of groups (Lackey 2021, 12).

[129] Lackey assumes that the evidence relevant to the proposition that *p* will subsume beliefs that bear on it, but she presents a way to build the disclosure of relevant beliefs directly into the condition in the following way: "Full disclosure of *the beliefs and evidence* […]" (Lackey 2021, 97).

Now that we have established that high belief uniformity can bolster a group's agency, as the group is better equipped in virtue of shared intentions to achieve their goals, let us explore the kind of goals we have in mind here. Namely, goals that are epistemically relevant to the group in question as they relate to assessing the intentions, agency, and rationality of the group (Kopec 2019, 539). But this does not conclude our examination as it raises the question of what kind of epistemic goals should be considered relevant? In some cases, our own concerns will determine the relevance, in others it might be based on what the group has officially or unofficially stated to be their goal, and thus it becomes a matter of how closely the goals that the group is working towards align with the group's stated goals (Kopec 2019, 539). Yet another way forward is to argue that the group's social context determines what epistemic goals are relevant (Kopec 2019, 539).

So, gatekeeping can lead to increased uniformity of beliefs of the respective group members, which in turn increase the coordination of the group. In addition, consider that belief uniformity also increases the efficiency of groups as the doxastic gap between members is smaller, so there are fewer instances where group members need to make epistemic concessions which lowers the amount of time and resources that need to be assigned to epistemic deliberation/voting[130] in order to reconcile the group member's doxastic beliefs to match.[131]

Two Problems of Gatekeeping

I introduce two problems that show how putting this kind of normative epistemic gatekeeping into practice makes groups especially vulnerable to certain kinds of problems that arise from the fact that people sometimes get things wrong, even when they try to be good epistemic agents. The two problems are as follows:

> POISONED WELL. The epistemic harm of a group member acquiring a false belief is exacerbated when gatekeeping is employed by their group.

[130] Which are the two main ways in which intragroup disagreements can be resolved, according to Broncano-Berrocal and Carter (2020).
[131] How closely they need to match is up for debate, Goldman (2014) suggests 60%, but other percentages and methods altogether have been proposed.

> KNOWLEDGE DEPRIVATION**.** The epistemic harm of a group member dismissing a true belief is exacerbated when gatekeeping is employed by their group.

How could the first problem come about? Think of a group of experts that employ gatekeeping as a mechanism for getting as many things right as they can for as cheap as they can. There is a robust gatekeeping process that incoming information must pass through before it is presented to the group. This allows members to rely on the quality control imposed by the gatekeeping process and when information has been approved, the members can lower their guard and rely predominantly on distribution norms, reducing the resources needed to spread the information across the group. However, if the gatekeeping process fails, i.e., a group member forms a false belief despite using acquisition norms in accordance with the gatekeeping process, then all group members will eventually be "poisoned" by the false testimony.

The second problem arises when the gatekeeping process fails in a different way. To begin with, consider what Sanford Goldberg calls *coverage*, which can be summarised as a believer's reliance on a testimonial source to be both reliably informed of domain-specific facts, and, when informed, reliably disposed to testify on the obtaining of those domain-specific facts (Goldberg 2021, 175). I argue that KNOWLEDGE DEPRIVATION stems from a lack of coverage-reliability. According to Goldberg, coverage depends on the "[…] completeness of testimony within one's epistemic community" (Goldberg 2021, 176). Insufficient epistemic coverage can result in the exclusion of relevant facts and evidence (Nguyen 2020, 143). An example of this can be seen when a group member erroneously judges knowledge to be false and subsequently refrains from forming the relevant belief.[132] In this case the group member does not share this purported falsehood among their peers and the knowledge is thus kept from the group.

Although there is a lot of knowledge out of reach at any given time, this kind of deprivation seems egregious, as it is knowledge that; 1) would likely have been believed by the deprived hearers if it was testified to them, 2) and without this kind of gatekeeping in place, the agents

[132] Note that although Nguyen finds bad coverage to be "an epistemic flaw of epistemic systems and networks, not of individuals." (Nguyen 2020, 143), we are not in disagreement, as a group that employs gatekeeping can be considered an epistemic system, even though its effects are felt by individuals.

would not rely on it, thus making them more likely to seek out such information themselves. Considering Lackey's (2021) conditions on justified group belief then it also becomes apparent that the problems of gatekeeping prevent groups from obtaining justification for their beliefs,[133] as condition (2), which stated that a group is justified in their belief if full disclosure of evidence (and beliefs) would not result in the total set of beliefs fail to make sufficiently probable that p, would not hold. There are at least two objections to knowledge deprivation being a real concern in instances of gatekeeping that can be pointed out here.

Firstly, it has been argued that there is too much information available to us, and it is therefore necessary to commit to a selection of it (Foer 2017). In some instances, this need for information selectiveness has increased the demand for information curators, such as social media companies and news aggregating services, who have found themselves in gatekeeping positions to decide what information is relevant to us (Simons 2022). Regardless of whether these entities should be the ones dictating how accessible any information is to us, this observation raises the question whether some kinds of knowledge deprivation truly suggest insufficient coverage or if it is instead a consequence of having full coverage. Knowledge deprivation might simply be a non-problematic result of having too much information available to us that cannot be practically processed.

Secondly, it does not look like the knowledge-deprivation problem is unique to groups prima facie and can rather be seen as one way a reductionist could criticise testimonial knowledge. Epistemic agents are often in a position to epistemically gatekeep others by refusing to share information with them. An even stronger way to do so would be to lie convincingly, in which case the hearer might be content with the false information they received and thus have less incentive than before the testimony to gather further information about the subject. By contrast, if they received no testimony, they might still feel motivated to inquire further.

We can respond to the first objection, that some kinds of knowledge deprivation are simply a natural consequence of having access to too much information and we have means to curate the

[133] At least it prevents groups from having justification for beliefs that have been affected by the gatekeeping problems, it is arguable whether the group has justification for other beliefs that have not been affected by the gatekeeping problems, but that is beyond the scope of this chapter.

information we need, by denying that this phenomenon counts as knowledge *deprivation,* and is instead a case of knowledge *selection.* That is, knowledge that is *withheld* differs from knowledge that is not selected, as the agent is denied access to information in the former case while the agent has access in the latter case. Even if we would concede that there is no practical difference between deprivation and selection, we can still point out that curating one's information takes epistemic work and some of the available information will be false. So more information does not necessarily result in an epistemically better body of belief.

Furthermore, it is important to note that when epistemic agents rely on others to curate information, the curators might have aims other than epistemic ones, such as the profit-driven goals of social media firms. Finally, we can concede that even if some characterisation of knowledge deprivation is unavoidable in cases of full coverage, making it necessary to commit to a particular selection of available information, the importance of *how* that selection is made remains just as significant. Conditions for justified gatekeeping can guide us in choosing which norms to employ to reduce the risk of a flawed selection of information.

To respond to the second objection, that the knowledge-deprivation problem is not unique to groups and can instead be viewed as a general reductionist critique of testimonial knowledge, we must distinguish between filtering as it is commonly used, and the specific type of gatekeeping used throughout this chapter. Although these terms have sometimes been used interchangeably, we can define filtering as a specific type of gatekeeping that happens at the hand of the speaker, rather than at the hand of the hearer.[134] Through this process, information is filtered by groups or individuals before reaching the recipient, ranging from individuals (e.g., when social media firms provide individually targeted content) to large audiences (e.g., when global news organizations report on something). Meanwhile, gatekeeping as presented here focuses on a kind of hearer-filtering done by a member of a group when they either dismiss or accept inter-group testimony and, if accepted, share the information intra-group.

We can now point out the difference between an individual filtering by deciding whether to withhold information or share it with a hearer, and members of a group following distribution

[134] Note that such filtering can be self-imposed in cases of self-selected informational networks. For further insight, see e.g., Pariser's (2011) *The Filter Bubble*.

norms. Namely, when groups have insufficient coverage as a result of KNOWLEDGE DEPRIVATION, then it is uniquely problematic because group members rely more heavily on distribution norms than individuals outside the group. Even if filtering can lead to individuals being knowledge deprived, they are still following mainly acquisition norms. The difference is that when a group member dismisses information, it affects every member of the group, whereas non-grouped individuals are affected only themselves. This makes knowledge deprivation uniquely and systematically problematic for groups compared with non-grouped individuals.

Nguyen (2020) defines *epistemic bubbles* as "a social epistemic structure which has inadequate coverage through a process of exclusion by omission" (143). We can think of gatekeeping groups that suffer from KNOWLEDGE DEPRIVATION as being instances of such epistemic bubbles. Note that POISONED WELL and KNOWLEDGE DEPRIVATION stem from *accidental* inclusion or exclusion of information. Malicious kinds of filtering, like we find in orchestrated *echo chambers*[135], will be problematic regardless of whether the two problems presented here can be accounted for.[136] For our purposes it is sufficient to say that KNOWLEDGE DEPRIVATION can generate epistemic bubbles, wherein relevant epistemic sources are simply left out rather than actively discredited (Nguyen 2020, 143).

Although Nguyen (2020) is not too concerned about epistemic bubbles, stating that they are rather easy to burst (153),[137] I would like to push back against this to underscore that knowledge deprivation as a result of group gatekeeping is a problem. Nguyen (2020) finds that to neutralise epistemic bubbles, one just needs exposure to information that is excluded from one's standard network, but still available (153). As it is an epistemic duty to gather relevant information proactively, epistemic bubbles are just a result of epistemic agents failing to perform their epistemic duties and could even be blamed for being epistemically lazy (Nguyen 2020, 153).

[135] Echo chambers and epistemic bubbles are both structures of exclusion, but only echo chambers exclude by manipulating trust and credence (Nguyen 2020, 141-142).
[136] The aim of this chapter is to show how gatekeeping with good intentions can introduce problems and what conditions should be in place to minimise them. The problems of malicious gatekeeping are beyond the scope of this chapter, but for a related discussion on information resistant agents in social networks, see chapter 18, "Learning from Ranters", by Morreau and Olsson (2022) in *Social Virtue Epistemology*.
[137] Nguyen (2020) states that "[e]pistemic bubbles are rather ramshackle – they go up easily, but they are easy to take down." (Nguyen 2020, 153).

While this approach can explain cases of filtering at the hand of the testifier, where the information is available, it fails to explain the challenge of groups afflicted by KNOWLEDGE DEPRIVATION. Members of such a group might not have the omitted information available to them because they trust other members to make good epistemic decisions, relying on acquisition norms when assessing inter-group testimony. If a member erroneously dismisses good information, then it is far from certain that other group members will encounter the same information at a later stage. In addition, it is difficult to claim that when a member believes they have gotten things right they should continue to inquire to verify that they are not in an epistemic bubble, as it becomes unclear at which point one would be allowed to accept what they are being told. If one does not know that one does not know then it seems harsh to place epistemic blame on them for failing to search for something they do not know is missing.

Recall that one of the potential effects of gatekeeping is increased belief uniformity, which can increase efficiency and coordination. Although this can be beneficial, it also means that if the gatekeeping fails, resulting in either POISONED WELL or KNOWLEDGE DEPRIVATION, the group is at risk of operating on a deficient body of knowledge in a coordinated and efficient manner. To illustrate this complication through a metaphor, imagine that someone finds a way to reduce the time needed to build a house, but through the same mechanism there is an increased chance that the blueprints are flawed.

One worry here is that a group with a defective total body of beliefs, as a result of the gatekeeping process failing, could still appear to have more total justification for its beliefs than a group with lower belief uniformity, since the latter would have more intra-group dissent.[138] To illustrate, a group having high belief uniformity suggests that a significant portion of the group members will have the same justification for their beliefs, making the group highly justified in its relevant false group belief. Lackey (2021) states that a group that is epistemically isolated and forced to obtain information from epistemically questionable sources, "[…] could end up more justified than one that engages in collective deliberation and forms beliefs on the basis of pooled evidence that has been scrutinized" (93). To be clear, we are not suggesting that gatekeeping as a function epistemically isolates groups, as members of gatekeeping groups are

[138] At least if we go by a Condorcet-inspired account of justified group belief (Lackey 2021, 91-95).

ceteris paribus not forced to turn towards certain sources and reject others. Additionally, the information that makes its way into the group will have been vetted by members following acquisition norms, so there is no reason to restrict access to information sources.

To recap, gatekeeping can lead to belief uniformity in addition to cheaper intra-group transmissions, and belief uniformity enhances group coordination because the group members increasingly share a specific body of knowledge. Unfortunately, gatekeeping failures can lead to POISONED WELL and KNOWLEDGE DEPRIVATION, which means that the groups that employ gatekeeping are at risk of not only operating on a false body of knowledge but doing so in a coordinated and efficient manner.

The Limitations of Gatekeeping

One attempt to resolve these problems is to assign more resources to the gatekeeping process. As Henderson (2011) writes: "[r]oughly, as one's stakes go up, it seems reasonable to be willing to pay higher "information costs" in order to guard against a wider range of failures" (88).

For simplicity's sake, think about a scenario in which ten individuals form a group. This group decides to employ a gatekeeping process to minimise the amount of epistemic work needed for knowledge transmissions. One way of implementing a gatekeeping process is to assign any single person to act as a mediator between the group and everyone else whenever new information is presented. This kind of gatekeeping process would support very cheap transmissions, as only one member out of ten would have to use acquisition norms when acquiring new information. However, this also means that the gatekeeping process is rather weak; the group is counting on a single person to get things right. If they get something wrong, the group will suffer from either of the gatekeeping problems. It is important to emphasise here that even good epistemic agents can err, and following acquisition norms does not *ensure* that

one avoids falsehoods and believes truths, even if the chances are lower than if one would follow distribution norms in the same setting.[139]

Now, let us imagine nine out of the ten members of the group decide to partake in the gatekeeping process. This means that whenever new information is presented to the group, nine out of ten members will review the information, following acquisition norms, and try to collectively figure out whether it should be believed. If the information gets through this relatively strict process, then the tenth member gets it for cheap.

Generally we find that strengthening a gatekeeping process is either done by group members shifting the balance even further towards acquisition norms when acquiring knowledge from outside their group, but it can also be done by increasing the number of people reviewing the incoming information. The point we are going to raise here is that incorporating distribution norms will remain problematic as long as acquisition norms cannot guarantee that one will avoid false beliefs, regardless of how the gatekeeping process is strengthened.

We can see that there is a relation between gatekeeping and the epistemic work being done by group members. If the resources used in the gatekeeping process are decreased/increased, then the gatekeeping process will have a higher/lower chance of letting falsehoods through or depriving the members not involved in the gatekeeping process of knowledge.

But, when we add additional resources to the gatekeeping process to decrease the chance of problems, the trust group members have in the gatekeeping process increases. We can see that any amount of gatekeeping induces less epistemic work than would otherwise be required, so if the gatekeeping process fails, there is a gap between the epistemic work required to meet acquisition norm standards and the epistemic work that is being carried out.

For a more detailed explanation, think about how testimony can serve as the acquisition, or the distribution, function of the concept of knowledge, and then consider that most testimony lies

[139] Note here that if gatekeeping would be done in such a way that completely eliminates the risk of the two problems, then the likelihood of epistemic misjudgements in cases of intra-group testimonial exchanges is the same regardless of the norms used.

somewhere between those two extremes. That is, most people will use both acquisition norms and distribution norms to various degrees when considering testimony. To demonstrate, consider the difference between a mother telling her son that there is milk in the fridge, a car salesman giving a customer information about a car that is for sale, and a police officer interrogating a suspect (Greco 2020, 4).[140] One way to look at what is happening in these cases is that the norms shift depending on the relation between speaker and hearer (and the subject matter). Now, in cases where the gatekeeping process is weak, the members of the group will use acquisition norms between themselves to a lesser extent than they would without gatekeeping, but they would still rely on it to a greater extent than they would if the gatekeeping process was strict. When the gatekeeping process is reinforced, the use of intra-group acquisition norms decreases, and so the false or dismissed beliefs that result from the gatekeeping problems will have a higher credence among the group members than they would otherwise have.

This means that simply increasing the quality of the gatekeeping process does not solve our two problems.[141] Another way to think about this is to see at which point the gatekeeping process would be completely reliable (or, as reliable as using acquisition norms exclusively). A fully reliable gatekeeping process requires either everyone to assume the role of a gatekeeper in all cases, or everyone would use acquisition norms exclusively within the group (effectively negating any benefits of gatekeeping as intra-group transmissions would never come cheaply). In either case, the epistemic work needed would be the same as it would be without a gatekeeping process. Recently, Dormandy and Grimley (2024) claim that one of the markers of gatekeeping success is whether it appropriately balances leniency and tightness, thus avoiding both *excessive leniency* and *excessive tightness* (393). This mirrors how an inappropriate balance between acquisition and distribution norms can lead to POISONED WELL and KNOWLEDGE DEPRIVATION.

[140] These cases are borrowed from Greco, who uses them, in part, to highlight how epistemic labour is divided differently between hearer and speaker depending on the situation.

[141] On a related note, Zollman (2013) finds that increased communication within groups of weakly connected individuals can be harmful when it comes at a cost without improving the epistemic performance of the group. Furthermore, even if the communication is free and non-redundant, it can still be harmful because it causes inquiry to be abandoned too early (25).

As we increase preventive measures to the gatekeeping process to decrease the likelihood of the two problems appearing, the epistemic labour savings go down. If we could decrease the epistemic harm of POISONED WELL and KNOWLEDGE DEPRIVATION, we could potentially justify gatekeeping.

I propose that gatekeeping can be justified for groups if they meet certain conditions. When these conditions hold, groups that aim to lower the epistemic costs of knowledge transmissions and increase their intra-group belief uniformity can proceed to gatekeep while having the risks minimised. After introducing the conditions for justified gatekeeping, we will show that the kinds of groups that best fit the criteria will generally be epistemically autonomous and composed of experts. This observation lines up with what we would expect, namely, that the groups less vulnerable to the two problems of gatekeeping are more likely to implement it.

Justified Gatekeeping

We have seen that POISONED WELL and KNOWLEDGE DEPRIVATION are not easily bypassed. As the gatekeeping process is strengthened to reduce the likelihood of the problems occurring (i.e., by increasingly relying on acquisition norms and thus proportionally reducing the reliance on distribution norms), the severity of them is potentially amplified.[142] Strengthening a gatekeeping process by shifting further from distribution norms to acquisition norms increases the total epistemic work done by the group. In lieu of traversing the scale between acquisition and distribution norms to find the optimal strength of gatekeeping that groups should aspire to implement, we can offer conditions for justified gatekeeping. The general aim of these conditions is to reduce the potential harm caused by the gatekeeping problems while avoiding an unnecessary increase in the total epistemic work done by the gatekeeping groups.

[142] Whether the lower risk of the problems occurring neutralises the increased fallout if they occur is difficult to determine. Modelling this could potentially give us parameters that show maximally effective gatekeeping in relation to risk, but they would depend on the composition of the group and how effective acquisition norms are at shielding us from false beliefs.

The conditions for justified gatekeeping rely on a notion of groups being capable of having beliefs and have meaningful disagreements. It has been argued that certain questions[143] about groups in an epistemic context cannot be answered using individual-based approaches. (Broncano-Berrocal and Carter 2020, 2). One such question pertains to the nature of group beliefs. There are at least three main accounts of group beliefs: summativists, non-summativists (or collectivists), and anti-realists. Summativist views are of particular interest here, as they suggest that the epistemic properties of groups cannot be wholly reduced to the epistemic properties of the groups' members (Broncano-Berrocal and Carter 2020, 2). Such groups, that constitute something greater than the sum of their parts, can be held responsible for their actions because their actions are attributable to the group as a whole,[144] rather than to individual members (Lewis-Martin 2022, 283). This is different from the non-summativist notion that groups are nothing more than the sum of their parts and as such their actions are reliant on the attitudes and behaviours of individual members, which Lewis-Martin (2022) states are only groups in a metaphorical sense (283).

The existence of metaphorical groups is generally reliant on shared intentions between the group members to interact with each other in a specific manner and perform (often necessarily repeatedly) various joint actions. [145] Groups, as they are understood from a summativist perspective, are themselves the source of the activity being performed.[146] As such, the group "[…] sets the agenda, regardless of the personal views of the singular agents who enable it" (Lewis-Martin 2022, 283). So, even though group members generally make up most of their group, their actions are limited by the group itself. If they were not constrained in their potential actions, then the group would not be the agent and its agency would be reduced to a metaphorical concept (Lewis-Martin 2022, 283).

What is evident here is that both summativists and non-summativists agree that groups are fit knowers (Kallestrup 2022), regardless of where the source of activity originates from, which

[143] For example, questions relating to epistemic adversity (Broncano-Berrocal and Carter 2020, 2).
[144] For further discussion on this point, see e.g., Lackey (2021).
[145] See e.g., Bratman (1992) and Gilbert (2009) for a general overview on shared intentionality, and Greco (2020) for a more detailed discussion on shared intentions as they relate to knowledge transmission.
[146] Note however that the non-summativist concept of groups does not necessarily make it eliminativist about group epistemic phenomena (D. Sosa 2023, 2).

suffices for what we are trying to accomplish here. However, David Sosa (2023) has argued for an anti-realist position which states that groups just feature a collective of beliefs and nothing more, so groups cannot have beliefs or tell lies, and that groups cannot assert on behalf of their members. Presumably, this has the implication that groups cannot disagree with each other in a meaningful way. Sosa's argument is largely based on the notion that belief requires consciousness (or at least the potential to be conscious), and that there is no way for individuals in a group "to come together to constitute a single locus of consciousness" (5). On his view, individuals express their rational nature when making rational judgments using their faculty of reason, something that groups cannot do. Groups can only engage in reasoning insofar as the reasoning is distributed amongst their members (D. Sosa 2023, 7). The group perspective is not singularly subjective but rather a collection of perspectives of all the members of the group and they cannot be unified into a common subjectivity in which individual rational deliberation would take place (D. Sosa 2023, 7).

It is sufficient here to claim that when members of a group believe something and the bases of their beliefs cohere, then something epistemically significant has happened. Even if this phenomenon is importantly different from rational states such as having beliefs, it still possesses the properties needed for us to move forward. That is, when a significant percentage of operative group members believe something and the bases of their relevant beliefs yield a coherent set of beliefs, and the group acts (either as an agent or just as the manifestation of a set of the beliefs of the group members) on those beliefs, then we can still pragmatically ascribe agency to the group even in the absence of consciousness.

We can now introduce conditions for justified gatekeeping that serve to minimise the risks of POISONED WELL and KNOWLEDGE DEPRIVATION. Groups are justified in their gatekeeping to the extent that they follow the following three conditions for justified gatekeeping:

(1) The group must be a part of a collective of groups that form an epistemically interrelated community.
(2) The group is sufficiently epistemically autonomous.

(3) The groups within the community must be at least subjectively epistemically rational, i.e., they have the goal of believing truths, avoiding falsehoods, and abide by epistemic practices that they believe are effective in achieving that goal (Mathiesen 2006, 165-166).[147]

First condition

The first condition establishes groups as the kind of agents these conditions are concerned with, as they are in an advantageous position compared with individuals and larger communities to gatekeep. It also stipulates the kind of communities these groups must belong to, namely communities that allow for communicative exchanges between groups. The idea here is that the focus should be on groups as an intermediate social level between individuals and larger communities that are composed of individuals and groups.[148] There are parallels between these three social levels that can illustrate why groups that fulfil this condition are ideal candidates for gatekeeping.

We can envision the larger community, composed of groups, as being itself a higher-order group which is not vulnerable to gatekeeping. To illustrate why, consider that POISONED WELL can occur within a gatekeeping group because its individual members are not contributing enough epistemic labour. In contrast, a community, understood as a higher-order group, is composed of groups that do not rely on distribution norms when communicating with one another, making it unlikely that the whole community will be affected.

Note that this is the case even if the groups within the community gatekeep by relying on distribution norms intra-group. With regards to KNOWLEDGE DEPRIVATION, we see that even if a group dismisses good testimony from another group or individual within the same

[147] This condition falls in line with what some social epistemologists have suggested to be a core facet of group rationality, see Kopec (2019). It is also worth noting that this teleological condition is not necessarily compatible with knowledge-oriented teleological accounts that are non-consequentialist, e.g., Neil (2016) and Littlejohn (2018).

[148] This distinction can be found in Rolin (2015, 163), albeit within a context of normative approaches to values in science.

community, the information is not deprived from the larger community as it remains with the speaker.

Second condition

The second condition, that each group is sufficiently epistemically autonomous, is imposed to mitigate the epistemic risk that is introduced by one group having too much direct influence over the beliefs of other groups. We can define epistemically autonomous groups as having their beliefs not *directly* influenced by other agents (Dellsén 2020, 349). [149]

According to Finnur Dellsén (2021), agreements reached by independent thinkers is more likely to be correct than an otherwise identical agreement reached by dependent ones (9).[150] This suggests that the larger community stands to benefit from the epistemic autonomy of its groups, as no single group commands the epistemic authority to directly shape the beliefs of others.

Groups being epistemically dependent on others is not inherently detrimental; the community can benefit from when a group, that can directly influence the beliefs of other groups, has gotten things right. The risk is that such groups are as vulnerable to POISONED WELL and KNOWLEDGE DEPRIVATION as the dependent groups, who would be risking epistemic harm by proxy.

One could argue that gatekeeping can reduce the epistemic autonomy of its members intra-group by promoting higher belief uniformity, thus conflicting with condition (2). However, it only requires the groups, as agents, to be epistemically autonomous, regardless of whether their members depend on others within their respective groups.

[149] Note that Dellsén makes a distinction between autonomy_A and autonomy_B, where autonomy_A has to do with acceptance rather than belief (which falls under autonomy_B). This is partly in response to Zagzebski's anti-egoistic argument, that privileging one's own belief-forming mechanisms would be incoherent (Zagzebski 2007, 257), by appealing to the notion of expert acceptance, but this distinction is not necessary for this chapter.

[150] The argument as it is presented here is incomplete, as even if unanimity in and of itself gives less reason than consensus to believe some theory, there might be other reasons to think that unanimity provides a greater reason to believe some theory. For further discussion on this point, see Dellsén (2021).

Another worry here is that the diversity of expertise is so great between groups that this condition will never be satisfied, as any group with a domain-specific expertise becomes an epistemic authority within their domain. This worry is exaggerated. The condition only requires that groups maintain *sufficiently* independent beliefs, and there is a difference between domain-specific epistemic dependence and general epistemic dependence. We do not want to claim that groups cannot rely on experts. A group having sufficient epistemic autonomy still allows for such testimonial exchanges as long as they can exercise acquisition norms appropriately without having their beliefs directly influenced.

Third condition

The third condition, which states that groups within the community must be at least subjectively epistemically rational. This means that groups must abide by epistemic practices that they believe are effective in achieving the goal of believing truths and avoiding falsehoods (Mathiesen 2006, 165-166). Without subjective epistemic rationality, there is a risk the groups would lack the epistemic motivation to self-correct.

An immediate objection here is that subjective epistemic rationality does not imply objective epistemic rationality. Nothing would prevent groups from being epistemically misguided, as long as they believe their practices effectively achieve their epistemic goals. Gatekeeping failures can thus lead a subjectively epistemically rational group to deviate from objectively rational epistemic goals in favour of other goals.

We can respond by conceding that this could be the case in the short-term. However, requiring groups to be subjectively epistemically rational minimises the long-term harms of gatekeeping failures. While justified gatekeeping may cause short-term epistemic harm, ensuring that the groups have truth-conducive goals helps mitigate these harms in the long run. Groups motivated by truth will, theoretically, converge on the best-supported theories as evidence accumulates.

The epistemic propriety of the methods employed by groups to come to an agreement (such as deciding what to believe) can be evaluated by their conduciveness to truth, and the reliability of those methods depend on both the individual and collective conditions in place (Broncano-

Berrocal and Carter 2020, 4). One might question how gatekeeping can be conducive towards truth if it opens up a possibility of POISONED WELL and KNOWLEDGE DEPRIVATION. However, it is important to establish what it means for a method to be conducive towards truth. There are important differences between individual agents and group agents. The aim is to demonstrate that the community can be more conducive towards truth through gatekeeping. This argument relies on a weak form of epistemic consequentialism, which would allow temporary epistemic setbacks for some groups if it enables the larger community to achieve truth more efficiently. It is weak because, although we are challenging the widely accepted idea that groups should aim at truth in order to be epistemically rational, we still assert that the epistemic community as a whole is aiming at truth, even if some groups are only subjectively doing so (Kopec 2019, 519).

There is a worry here that, testimony from the affected groups gets through the gatekeeping processes of other groups if we allow these short-term epistemic gatekeeping failures to occur, causing further groups to adopt false beliefs. This would be a pessimistic view of the gatekeeping process. Gatekeeping groups will be following mainly acquisition norms in inter-group testimonial exchanges. The groups would therefore not be in a more epistemically vulnerable position than they would assessing any other false testimony.

A stronger version of this objection considers whether the rate of gatekeeping failures could outpace corrections, leading to widespread misinformation. If this were to happen, it would simply point towards some imbalance in the allocation of acquisition and distribution norms within the groups. One way to address this objection is to make sure that the groups within the community assign *enough* resources to their gatekeeping, which would lower the risk of POISONED WELL and KNOWLEDGE DEPRIVATION enough to prevent this vicious chain reaction to a reasonable degree.[151] The collective effort of the community as a whole can thus reduce the likelihood of such failures.

[151] Note that even if this manoeuvre results in more epistemic labour, it is still cheaper than if members of the groups were to use strictly acquisition norms in their epistemic endeavours.

Gatekeeping Eligibility

With these conditions in place, what kinds of groups exist that fulfil them? I argue that groups that are a part of an expert community, such as a scientific community composed of research groups, are in an advantageous position to justify gatekeeping. This is an unsurprising conclusion. Dormandy and Grimley (2024) suggest that science is an "epistemic gold standard", where the term indicates epistemic quality and certain ideals of inquiry (392-393). They further argue that gatekeeping in science seeks to preserve those standards (Dormandy and Grimley 2024, 392-393).[152] We have generally been focused on gatekeeping as it relates to knowledge generally, but as we shift our attention to practical applications of justified gatekeeping, such as in science, we see that it aims to exclude bad science and facilitate knowledge sharing. Additionally, Greco (2020) finds that there are various institutional and social practices in science that are aimed at bringing high-quality information to the relevant scientific community (governed by acquisition norms) as well as distributing that information within that community (governed by distribution norms) (40).

Dormandy and Grimley (2024) further suggest that the success of gatekeeping processes in science should be evaluated based on whether it preserves what it seeks to preserve without including things that should be kept out (which could lead to POISONED WELL), or excluding things that should be let in (which could lead to KNOWLEDGE DEPRIVATION) (393). The way these problems can appear in practice is for example when background beliefs from the surrounding culture infiltrate the scientific framework which results in POISONED WELL, and when a dissenter is excluded despite pointing out genuine flaws in the scientific framework which results in KNOWLEDGE DEPRIVATION (Dormandy and Grimley 2024, 394).

The ever-increasing specialization within the sciences, along with limited time and resources, makes it rare that any one group member is in a position to acquire all the relevant knowledge required for any particular research project (Barimah 2024, 3). As a result, collaboration and the distribution of knowledge have become essential components of scientific progress

[152] Dormandy and Grimley (2024) note that these are just characterizations of some of the important features of science, and not an attempted definition of it (2024, 393).

(Barimah 2024, 3).[153] Although expert communities, as a collective of groups, often have a loosely defined collaborative environment in place, the inter-group interactions of those groups do not maintain the same level of distribution norms as they do intra-group. Groups within expert communities are thus in a strong position to implement justified epistemic gatekeeping.

Conclusion

Groups that implement epistemic gatekeeping to access cheap knowledge transmissions and, in some cases, belief uniformity, are epistemically vulnerable to the two problems of gatekeeping. Furthermore, these gatekeeping problems, POISONED WELL and KNOWLEDGE DEPRIVATION, are difficult to manage because in order for groups to lower the risk of gatekeeping failures they must increase the epistemic labour involved by shifting further towards acquisition norms. I proposed three conditions for justified gatekeeping that aim to minimise the risks involved. The conditions for justified epistemic gatekeeping are; (1) the group must be a part of a collective of groups that form an epistemically interrelated community, (2) the group is sufficiently epistemically autonomous, and (3) the groups within the community must be at least subjectively epistemically rational, that is, they have the goal of believing truths, avoiding falsehoods, and abide by epistemic practices that they believe are effective in achieving that goal.

These conditions are most naturally fulfilled by expert communities in virtue of the kind of group structures found in those kinds of collectives. Furthermore, they generally possess shared truth-conducive goals that encourages them to correct their course in case of gatekeeping failures. Gatekeeping involves risks that need to be taken seriously, and having conditions for justified gatekeeping can mitigate those risks.[154]

[153] For example, Rolin (2015) and Hardwig (1991) find that trusting colleagues (group members) for information one does not possess is important for epistemic success, although Frost-Arnold (2013) acknowledges that such trust might be motivated by self-interest, i.e., the tarnishing of one's reputation in case of fraudulent work and acknowledgment in case of exceptional work.

[154] This aligns neatly with what Goldberg (2021) concludes about the epistemology of coverage in *Foundations & Applications of Social Epistemology*, namely, that the epistemology of coverage is about assessing how well individuals and communities manage various risks as they aim to reap the benefits of their information-saturated environment (Goldberg 2021, 189).

Chapter 5: Artificial Competence

Abstract. This chapter introduces a virtue-theoretic distinction between general AI and narrow AI systems by analysing AI competence through Ernest Sosa's (2021) virtue-theoretic framework. First, a general distinction is made between narrow artificial intelligence and general artificial intelligence. Then, I characterise AI in a way that bridges its terminology with virtue epistemological concepts, establishing a foundation for meaningful parallels between the two. After that, I introduce the virtue-theoretic framework that will be employed to analyse AI competence. I then show how a clear distinction can be made between narrow AI and general AI based on whether an AI system is capable of reflecting on their predictions using second-order competence. I find that increasing the competence of AI in a specific domain is not sufficient for it to be considered a general AI, and that constitutional competence in an epistemic domain is a necessary quality of general AI. Finally, I suggest that this virtue-theoretic distinction between narrow and general AI can meaningfully contribute to other areas of philosophical research, for example on the nature of intentional behaviour and in the evaluation of AI trustworthiness.

Artificial intelligence

Artificial intelligence technology (AI) has advanced rapidly in the last few years. Recent innovations in both hardware and software factors contribute to this sudden shift of pace in AI development. The most important ones being a dramatic increase in processing power using specialised computational hardware (such as CPUs, GPUs, TPUs[155] and even early-stage qubit computing) (Zhu, et al. 2023), and groundbreaking discoveries in machine learning using technological neural networks (OpenAI 2023). This rapid development of AI has resulted in unprecedented breakthroughs in image processing and recognition (OpenAI 2024), natural language processing (OpenAI 2023), audio manipulation (OpenAI 2024), and autonomous systems more generally (OpenAI, Hello GPT-4o 2024).

[155] See e.g., Sato (2018) for further information on TPU processing.

Large language models (LLMs), such as ChatGPT (Generative Pre-trained Transformer), are computational devices used for natural language processing and are able to generate human-like text and complete other language-related tasks with high accuracy (Kasneci, et al. 2023, 1). They are designed to generate sequences of words, code, or other data, from a source input (prompt) (Floridi and Chiriatti 2020). ChatGPT-4o[156], a modern LLM made by OpenAI, embodies both the fantastical advancements of AI, as well as its inherent limitations. While ChatGPT-4o is able to instantly generate coherent and contextually relevant responses, easily passing the Turing test (Mei et al., 2024), it also generates perplexing non-sequiturs, tells brazen lies, and misunderstands the simplest tasks (Hicks, Humphries and Slater 2024).

Furthermore, even when ChatGPT and other generative AI LLMs produce coherent and accurate text they are often unable to provide evidence for the claims they make. Instead, they make up sources and facts that do not exist. This tendency of LLMs to fabricate evidence has been called the problem of "AI hallucination" (Hicks, Humphries and Slater 2024, 38). LLMs like ChatGPT do not reflect on the text they produce, which can make these hallucinations (or confabulations[157]) snowball[158], generating further errors (Zhang et al., p. 1). Interestingly, in some cases, when LLMs are asked to justify their previous hallucinations, they generate false claims that they can recognise as being incorrect when they are presented with the same false claims in a separate interaction session (Zhang et al., 2023, p. 2). Recently, Hicks, Humphries, and Slater (2024) have argued that ChatGPT is a *bullshit machine*, and instead of hallucinations, ChatGPT's erroneous responses should be called *bullshit* because the terms "hallucinations" and "confabulations" suggest that ChatGPT can perceive (if it can hallucinate), or rely on memory in a traditional sense (if it can confabulate) (Hicks, Humphries and Slater 2024, 8). Furthermore, both terms suggest that ChatGPT is generally attempting to convey accurate information, when it is simply predicting the next word in a sentence.

The gap between human cognition and machine intelligence has diminished *prima facie*, but the remaining contrast has become even starker. Recent developments in generative AI have

[156] A fourth-generation autoregressive language model that employs deep learning on large internet datasets made up of texts to produce text that could pass as being written by humans (Cassinadri 2024, 1-2).
[157] As suggested by Edwards (2023).
[158] Zhang et al. (2023) refer to this phenomenon as *hallucination snowballing* (1).

forced us to reevaluate the difference between human cognition and machine intelligence, and furthermore, raise the question of what requirements should be considered for generative artificial intelligence, such as LLMs, to be thought of as real intelligence?

This chapter attempts to answer this by introducing a virtue-theoretic epistemic explanation of AI competence and AI knowledge, which allows us to have clear boundaries between narrow AI and general AI, and furthermore, provides the requirements AI would need to fulfil to be considered a general AI. Ernest Sosa's virtue-theoretic framework is especially well suited to answer these questions because it takes competences to be special cases of dispositions, where competences are dispositions of an agent to perform well in a given domain. This allows us to use Sosa's framework *mutatis mutandis*, as AI can have dispositions regardless of consciousness or biology.

Before doing so, we must introduce some core concepts of AI that will be of use throughout the following sections. A good starting point to distinguish between the kinds of AI that will be the focal point of this paper, general AI and narrow AI, is a thought experiment called *the Chinese room*, proposed by John Searle (1980). In Searle's thought experiment a native English speaker is locked in a room that is filled with boxes of Chinese symbols and instructions that tell them how to decide on which symbol should be used when presented with Chinese symbols that are sent in to their room. They are then given a series of Chinese symbols that are, unbeknownst to our English speaker, questions. By following the instructions, the English speaker is able to answer the questions correctly by relying solely on the instructions. The English speaker is thus able to pass the Turing test for understanding Chinese without understanding a single word of Chinese (Searle 1999, 115). Searle's point here is that even if it appears that a program can understand something, it does not make it certain that it does so. Rather, it could just be manipulating symbols in a way that imitates understanding without actual understanding, and simulation is not the same as duplication (Searle 1999, 115).

Before moving forward, it will be beneficial to define narrow (sometimes called weak[159]) and general (sometimes called strong) AI. Narrow AI commonly encompasses AI systems that are built for specific tasks or applications that are well defined. They execute precise functions within a limited domain that cannot be generalised to tasks beyond that domain while general AI is commonly understood as having fully developed human cognitive capabilities (Sheikh, Prins and Schrijvers 2023). In this paper, an underlying notion of AI found in Searle's Chinese room, namely, that general AI would "have a mind in exactly the same sense human beings have minds" (Dennett 1991, 435). The difference can be understood as the difference between simulating a mind and having a mind. That is, narrow AI would simulate *a model* of the mind, while general AI would simply simulate *a mind*. We can now define general AI and narrow AI in a way that is representative of how these terms are commonly understood:

> **Narrow Artificial Intelligence (NAI)**: Specialised AI systems engineered to perform well-defined tasks within a limited domain or a set of domains. These systems are often capable of surpassing human performance within the restricted domain, but incapable of generalising their performance beyond their designated functions (Goertzel and Pennachin 2007).
>
> **General Artificial Intelligence (GAI)**: AI systems with human-like cognitive abilities, enabling them to gather sensory inputs[160], reason, learn, and adapt across a diverse range of domains (Goertzel and Pennachin 2007).

The distinction between NAI and GAI mostly relates to the range of domains the AI is proficient in. In short, NAI can be thought of as a task-specific optimised system in its relevant domain, while GAI can be said to display general cognitive autonomy as one would expect from normal epistemic agents (humans). GAI, as of writing this, does not exist. NAI, however, is rapidly integrating itself into our lives, and while image and audio generative AI systems have gotten increasingly more attention, LLMs, such as ChatGPT-4o, have become almost synonymous

[159] Sheikh et al. (2023) prefer using the term "narrow AI" instead of "weak AI" because the latter implies that such AI lacks strength, which is not necessarily the case. In contrast, "narrow AI" suggests that it is limited to a well-defined (or narrow) domain, which is more accurate.

[160] It is difficult to attribute a stronger notion (such as perception) to AI here, as that might be viewed as requiring phenomenal consciousness, which is not a claim this chapter will argue.

with AI. It is currently the flagship of LLM technology (OpenAI, Hello GPT-4o 2024), and will thus be used the standard example of LLM in this paper, although any sufficiently evolved LLM could be substituted *ceteris paribus*.

In this chapter it is important to highlight that even if we will be comparing AI attempts at a performance with human attempts to a degree, ChatGPT-4o only indirectly and incidentally tracks truths, as the design function of an LLM is to produce text that sounds plausible. However, it is not clear what follows from this concession. Consider, for example, that the relevant design function of paradigmatic epistemic competences like perception and memory is to help survival and more generally evolutionary fitness of the organism. This is not an epistemic aim, but an evolutionary adaptive aim keyed to organism fitness. Granting this fact, however, does not stand in tension with the idea that our perceptual and memory functions reliably deliver accurate information. The compatibility of epistemic faculties having non-epistemic design or etiological functions, thus non-epistemic function-generated aims, offer a vantage point to reassess what it means for LLMs to be optimised for producing plausible-sounding content. If that is their design function, they might have a function-generated aim that is not epistemic, but that in and of itself does not necessarily preclude them from having the kind of reliable connection with truth (a question determined by seeing how well they actually deliver true information) that is demanded by competence, at least in appropriate conditions.[161] For instance, if we specify that the kind of situational component of competence that is normal for LLMs to operate includes largely reliable training data, then it looks like we can make sense of a virtue-theoretic competence structure[162] forming the basis of what looks like reliable artificial competence, even if we continue to grant that LLMs are intentionally designed to produce plausible human-like text.

A different issue that must be addressed when trying to examine AI from a virtue-epistemological perspective is that it is generally thought incapable of possessing beliefs, instead settling for predictions using data they have been granted access along with self-generated data. However, not everyone agrees. Herman Cappelen and Josh Dever (2021), in

[161] For further clarification on the distinction between design functions and etiological functions as they relate to AI, see Simion and Kelp (2023).
[162] This is in reference to Sosa's SSS competence, which will be examined further in the next section.

their book *Making AI Intelligible*, have developed a de-anthropocentrised externalism, that, along with the extended mind thesis, suggests that AI can make statements that predicate properties. A very brief summary of their proposal is that a mental act cannot be the act of predication in isolation from function, and the relevant function is to give rise to judgements that guide our actions (Cappelen and Dever 2021, 123).

Regardless of whether AI can have beliefs in the traditional sense we are still in a position to claim AI competence, even within the epistemic domain, as long as we assume that LLMs are at least configured and optimised to make accurate[163] predictions (Landgrebe and Smith 2023). Consider that in unsafe modally close worlds one can retain epistemic competence regardless of safety. Whether we are in a matrix or not, our epistemic practices look the same.

If we at least claim that LLMs are making attempts at a performance, where the performance in question is something along the lines of "making accurate predictions according to the data available", then we can see that LLMs are in some sense reminiscent of the case of Norman the clairvoyant, who possesses perfectly reliable clairvoyance faculty (BonJour 1980). Norman poses a challenge for reliabilists, as he is not justified in his true beliefs even when his belief-forming process is perfectly reliable. The difference between Norman and AI is that although they are both making predictions, only the AI can *justify* its predictions with evidence and reasoning on account of the data it holds and the neural network processing it performs.

Note that the inner workings of AI can be akin to "black boxes", so their justification for what led to a given prediction might not be interpretable by humans. Consider that there might be thousands of variables that contribute significantly to a single prediction. Even if they could all be inspected and the way the AI system decided to weigh each variable was accessible[164], it is not reasonable to expect users to understand why the AI system came to a prediction (Ribeiro,

[163] Where the sense of accuracy is contingent on the version of the world that is simulated in the AI's training data (Landgrebe and Smith 2023).

[164] These kinds of AI systems have been called "white boxes", and although they have an epistemological advantage over black-box models, "they do not automatically yield solutions that facilitate responsible and accountable use in practice" because "complex white-box models are also opaque to most domain- and non-experts." (Herzog 2022, 223).

Singh and Guestrin 2016, 1137). Fortunately, AI can be imbued with explicability[165], which entails explainability and interpretability, often referred to as explainable AI, or XAI. This enables humans to retain intellectual oversight, generally by providing access to the reasoning behind the decisions and predictions made by the XAI, in the form of comprehensible explanations[166] that use concepts that we understand (Cappelen and Dever 2021, 25, Longo, et al. 2024). The opaque nature of AI systems makes XAI development focused on the explainability and interpretability[167] of AI vitally important to our understanding of present and future AI systems (Cappelen and Dever 2021, 26).[168]

In general, it seems that the literature on AI mirrors epistemology in some important ways. When AI make prediction attempts, they are *guessing*. Sosa (2021) gives an example of an eye-exam, where one starts to incorporate guessing as the letters get smaller. In his case, the guesses are correct, and reliably so (E. Sosa 2021, 144). However, even if one is trying to guess correctly, there is something preventing the guesses from being considered judgements —viz., the guesses might be apt (correct because of manifested competence), but they are not aptly apt (where the guesser must attain aptly not only the truth of their affirmation but also its aptness).

Furthermore, the literature on AI is concerned with performances and domains, which seem to fit well within a virtue epistemological framework such as the one Sosa has developed, most recently in his book *Epistemic Explanations* (2021). Additionally, Sosa's approach to competences takes them to be special cases of dispositions, i.e., dispositions of agents to perform well in a given domain. Thinking about competences as special cases of dispositions is advantageous to us when writing about artificial intelligence, as machines can have those kinds of dispositions regardless of knowing whether they can possess consciousness or the kind of biological architecture on which human cognition supervenes. Before we consider AI

[165] Explicability is a richer notion than explainability, as it combines intelligibility and accountability (Herzog 2022, 219, Floridi, Cowls, et al. 2018).
[166] Although as Ribeiro notes, "it is often impossible for an explanation to be completely faithful unless it is the complete description of the model itself", an explanation can still be meaningful if it is locally faithful, i.e., the explanation corresponds to how the model behaves in the proximity of the predicted instance (Ribeiro, Singh and Guestrin 2016, 1137).
[167] Christian Herzog has argued that there is a need to focus on explicability because neither explainability nor interpretability automatically incur accountability (Herzog 2022, 223).
[168] XAI development is still plagued by implementation difficulties in extracting explanations of AI system behaviour, and furthermore, explicability and interpretability are not clearly defined, and it is not clear what kinds of tools are required for interpretability (Cappelen and Dever 2021, 26).

competence from a virtue theoretical perspective, we need to elaborate on some of the key ideas found in Sosa's framework.

Virtue-Theoretic Competence

Sosa's telic virtue epistemological view argues that epistemic normativity should be understood as one form of telic normativity and that performances can be understood as attempts aimed at some goal. An attempt can thus become an achievement when the attempt is successful because of sufficient competence (E. Sosa 2021, 18). As we have already established, competences are a special kind of dispositions, namely, dispositions of agents to perform well in a given domain. With this in mind, there is no barrier to extending the basic framework Sosa has developed to artificial intelligence. Instead of focusing on the dispositions of human agents, we will be focusing on the dispositions of machines. To assess whether a performance (of a man or a machine) is good, Sosa's AAA structure is appropriate:

> **AAA**: Performances are Apt if and only if they are Accurate (they attain success) because they are Adroit, and they are adroit iff the performance's success manifests a complete competence (E. Sosa 2021, 18).

Aptness is a necessary and sufficient part of achievements, as without it one can be accurate and adroit because of luck, allowing for the possibility of lucky successes such as Gettier cases constituting as cases of knowledge. An oft-cited example to grasp these distinctions is of an archer. The archer shoots at a target and their shot hits the target. The archer's performance is accurate because they hit the target. It is apt iff the archer hit the target because their performance was adroit, i.e., the archer manifested their competence when they hit the target.

An important question is what it means to manifest competence. It will be beneficial moving forward to take a close look at the role competence plays in Sosa's AAA structure. In the archer example, the performance can be said to be completely competent if the shot manifested the archer's intrinsic archery skill (constitutional competence[169]), the archer was in good shape

[169] Sosa also refers to this as one's innermost competence (E. Sosa 2021, 46).

when shooting (inner competence), and the situation when shooting was appropriate (E. Sosa 2010, 465). Competences of an agent can thus be seen as dispositions they have to perform well and is comprised of the agent's skill, shape, and situation (E. Sosa 2010, 465). It is clear then, that competence is a vital element of fully apt knowledge. Consider the following definition:

> **SSS**: Competence is manifested by the intrinsic skill situated within the agent (seat of the disposition), the agent's shape as it pertains to their current ability to exercise their skill (shape), and the appropriateness of the situation they are in with regards to how it affects their execution (situation) (E. Sosa 2010, 465).

To further clarify, if the archer is to perform competently, they would first of all have to have the skills required to hit their target reliably enough when they are in proper shape and properly situated, as if they would not possess the necessary skills to do so they are incapable of competently hitting the target as luck would be the salient reason for them hitting it.

Secondly, they would have to be in a position to access those skills, i.e., the agent must be in shape. The details of what exactly constitutes good shape can vary in practice, but one can imagine that in the archer's case, being in shape means being sober enough to utilise their skills, keep their eyes open, be mentally unperturbed, alert, and so on. Note that these requirements depend on the notion that without them the archer could not perform up to the intrinsic skill they possess, e.g. if they could reliably shoot their target while drunk with as much skill as if they were sober, then the state of being drunk would be compatible with the shape needed to be competent.

Thirdly and finally, the situation must be appropriate for the archer to perform. Here, appropriate can be understood as favourable, that is, the situation must not be such that it intervenes with the attempt to such an extent as to make the archer fail their shot, given that the archer is in good shape and possesses the skills required otherwise. In this specific case, the appropriate situation could for example pertain to the winds being manageable, and enough daylight to see the target.

Now that we have explained what an apt performance entails, we must make a further distinction between apt performance, and aptly apt performance. Even when a performance is apt, in the sense that the agent is accurate because adroit, it is not necessarily aptly apt. For a performance to be aptly apt, the agent must be attempting not only to be accurate but apt. In other words, the performance itself is not just apt, it is aptly apt. This becomes an especially meaningful distinction when we move from athletic performances to intellectual performances, such as acquiring and possessing knowledge.

When we apply the AAA structure to epistemic concepts such as knowledge, we find certain similarities between the epistemic agent and the archer. On Sosa's view, beliefs are epistemic performances that aim at truth. Knowledge can thus be understood within the AAA structure as an apt belief. To illustrate, an epistemic agent that aims at truth is successful when they believe the truth (the belief is accurate), they formed their belief competently (the belief is adroit), and they believe the truth because of their competence (the belief is apt). Once again, we use the SSS structure to see what competence in this example entails. First, the epistemic agent must possess the relevant constitutional epistemic skills, such as cognitive abilities, to believe the truth (seat). Second, they must be in a good epistemic state to apply those skills, e.g. by being sober, awake, and alert (shape). Thirdly, the situation must not be epistemically hostile, e.g. the epistemic agent is not being deceived by tricky lighting or illusions, and they can utilise their epistemic skills without external interference (situation).

Sosa divides the various epistemic competences available to into two main categories; non-global competences, which host seemings, and global epistemic competences, which are basic judgment-forming competences (E. Sosa 2021, 148). Non-global competences can be further divided into three subcategories, derived competences, underived modular competences, and basic central-processing competences. Derived competences are attained through more basic competences, e.g. when an agent learns how a thermometer works through testimony or tests independent of the thermometer's readings, which results in the agent becoming competent in telling the temperature (E. Sosa 2021, 148). Underived modular competences host seemings with propositional content without relying on more basic competences, e.g., perception (E. Sosa 2021, 148). Basic central-processing competences reliably enough produce all-things-considered seemings, whether they are occurrent or implicit (E. Sosa 2021, 148).

Global epistemic competences differ from non-global competences in the sense that they are used in determining how to judge, all things considered. This means that nearly anything could be relevant to any epistemic attempt (given proper stage-setting). Sosa claims that this kind of "holistic competence is global in that it is required in properly making *any* judgment or forming *any* belief (E. Sosa 2021, 148).

These global and non-global competences are relevant to AI because they enable us to think more clearly about exactly what kind of epistemic competences LLMs possess. For example, consider that for an AI to possess the sort of holistic global epistemic competence Sosa is talking about, it would need to be able to make judgments *all things considered*. Conversely, LLMs seem to operate on derived competences to a considerable degree. Global and non-global competences will be revisited in the next chapter where they will be utilised to support a distinction between NAI and GAI.

This distinction between global and non-global competences is reminiscent of a distinction that can be found when considering different kinds of XAI explanations, which can be categorised as either being global or local. Global explanations are best understood as being two-place explanations while local explanations are three-place explanations. That is, global explanations can be generally defined as explaining AI predictions, while local explanations can be defined as explaining specific predictions of an AI. Consider the following definitions:

> **Local AI explanations:** Explanations of individual predictions made by a machine learning model, with a focus on understanding why a particular prediction was made.
>
> **Global AI explanations**: Explanations of how a machine learning model operates broadly speaking, with a focus on understanding the overall performance of an AI model.

To finalise our discussion of the virtue-theoretic framework[170] that will be used to analyse AI competence, we need to categorise the kinds of knowledge that are derived from this

[170] In *Epistemic Explanations*, Sosa (2021) expounds further upon this structure, distinguishing between alethic affirmation and judgement, causing this sort of animal knowledge to be redefined as apt alethic affirmation.

framework. We know that apt belief is knowledge, but what kind of knowledge? According to Sosa, it is knowledge that manifests complete competence in accordance with the SSS-model, but without requiring the constitutional competence (seat), the inner competence (shape), or external circumstances (situation) to be safe. Sosa calls this kind of knowledge *animal knowledge*.[171]

For knowledge to be safe, a further requirement needs to hold. Sosa calls this reflective knowledge, i.e., the belief's aptness itself manifests a second-order competence of the epistemic agent. Reflective knowledge is thus knowledge with the stipulation that one needs to have competently assessed whether one's first-order competence and external conditions positioned such that one's first-order belief is unlikely to have been inapt (Carter 2018, 285).

Artificial Competence

Let us now shift our attention back to AI and LLM. If we were to evaluate the performance of these models, how would we go about doing so? A good starting point is to outline what, if any, competences these LLMs possess.

One of the main differences between epistemic agents and LLM is that LLM can only predict and not know, indicating that the way we treat AI in epistemology is akin to a person that can only predict things. However, at a certain point AI is advanced enough that the predictions are more like Norman the clairvoyant. We have no understanding of why AI predicts that *p*, but we can be rather certain that *p*. Process reliabilists are thus faced with a new version of the clairvoyance problem in the form of a black-box AI that seems to be justified in their predictions as they have a high truth ratio, without any evidence or reasons about the accuracy of its predictions.

Recall the GAI/NAI distinction and consider once more their defining characteristics, namely, that NAI are specialised AI systems that perform tasks within a limited domain or a set of

[171] Introducing Sosa's full framework, including knowledge full well, secure knowledge, and background conditions, is unnecessary for the purpose of this paper, all that is required to grasp the competence of AI as it stands are the ideas on competence and how they relate to apt beliefs, and aptly apt beliefs.

domains while GAI have human-like cognitive capabilities across a diverse range of domains. If we are to distinguish between NAI and GAI using a virtue-theoretic framework, a plausible starting point would be to claim that NAI performance attempts to predict are simply not apt while GAI would be apt in their predictions. This would be too hasty. See the following example from Sosa's *Epistemic Explanations* about Simone the fighter pilot, who could easily not be in a real cockpit but in a near-perfect[172] simulation:

> "In my thought experiment, trainees are strapped down asleep in their cockpits, and only then awakened. Let us suppose Simone to be in a real cockpit, flying a real plane, and shooting targets accurately. Surely her shots can then be not only accurate, but also competent, and even apt" (E. Sosa 2010, 468).

It is clear that Simone is not properly situated from a second-order perspective in this example, and her competence is thus not manifested in a way that is apt. But, as Sosa points out, "what of her intellectual shots, her judgments and beliefs?" (E. Sosa 2010, 468). Now let us imagine that Simone forms the belief that she successfully shot a target, and her belief is accurate and competent. Sosa asks whether Simone's belief here can be apt, as well as being accurate and competent, in light of the lack of safety of her belief as she could very well have been placed into a simulation without knowing it (E. Sosa 2010, 468). Regardless of whether Simone could have been in a simulation, it seems that her shots accurately hit the target because of her shooting competence, i.e., Simone's "competence manifests in the accuracy of her shot" (E. Sosa 2010, 468).[173] When we apply this line of thinking to the epistemic domain, we can see

[172] Imagine the simulation to be completely akin to the real world except for whatever property distinguishes the real world from a simulation, i.e. there are "no tell-tale signs" that Simone is in a simulation (E. Sosa 2010, 468).

[173] In the case of narrow AI, such as LLM, we can see something similar happening. In that case we can think of LLMs as being in a kind of simulation, that is, the LLM has been trained on a dataset that cannot fully represent the world, but the LLM has no point of reference to realise that. One objection is to claim that LLM frequently seem aware of their own limitations, for example, when ChatGPT is asked about something that happened later than January 2022, it gives the following disclaimer: "My training data includes information up until January 2022. Anything that has occurred or been published after that date wouldn't be directly accessible to me unless it has been shared with me in this conversation." (OpenAI, ChatGPT 2024). However, consider that this only means that it has been given this information by those who control the data it is being trained on, in much the same way, if someone told Simone that she was in fact in a simulation, she would form the belief that she was in a simulation, but it would not help her identify whether the target she shot accurately would have been accurate if she was not in a simulation. So, LLMs, even when they know they are being trained on a limited dataset, cannot adequately account for the limitation they have been made aware of. This will be discussed further in this chapter in relation to non-global competences.

how beliefs might be apt without constituting reflective knowledge, as knowledge is generally considered incompatible with the sort of accidental luck portrayed in Simone's case.

We must now clarify whether AI is capable of virtue-theoretic competence, and if so, what kind of competence? Furthermore, if the kinds of competence that differ between narrow AI and general AI, can that be used to sharpen how these categories are defined? We start by examining the relevant domains, then we consider what constitutes an AI attempt, and what they are attempting to do. Next, we reintroduce the SSS-model with AI in mind, and then we present the types of AI competence from a virtue-theoretic perspective and demonstrate how this classification can aid us moving forward.

We have already established that competences are dispositions of an epistemic agent to perform well. These competences are not inherently epistemic, as we can clearly see when considering the various performance domains in which competence appears, such as sports, politics, science, professions, morality, and artistic domains. In addition, the notion of having competence within a domain depends on what the respective aim of the domain is. That is, competence is the agential disposition to attain the aim (and attain it aptly) of a domain (E. Sosa 2021, 45).

We have so far mostly been focused on general performance domains and competences, such as the case of the archer, because Sosa's telic virtue-theory can account for more than just epistemic instances. Now, let us redirect our focus to epistemic domains and competences, particularly as they pertain to cases of knowledge. As we are not only concerned with AI competence, but the differences between NAI and GAI, it does not suffice to simply determine whether NAI can have competence and if so, how it appears (e.g., image generation, text manipulation, and recognition capabilities). Focusing on these epistemic domains will be instructive as we begin to examine whether there are epistemic differences between NAI and GAI, and what epistemic properties, if any, are causally linked to those differences.

As we have seen, even though LLMs are often portrayed as sources of information, the aim of their attempts is only to make good predictions. Unfortunately, the goodness of AI predictions is only contingently connected to truths. Just as an AI can predict a truthful proposition, it can

predict bullshit. However, the requirements for reflective knowledge are inherently epistemic, regardless of whether the first-order competence is situated in an epistemic domain. For further clarification, note that a baseball player that is attempting to perform in the domain of baseball is not aiming at truth, but the epistemic domain is still relevant when they consider whether they are triple-S competent, as epistemic reflection is needed for this kind of a second-order competence. As we have covered previously, the competence of AI can be understood in a similar way as human competence, i.e., even if AI is not aiming at truth, the epistemic competence needed to retain full triple-S competence appears to be the same.

Now, if we tried to analyse AI competences with regards to non- epistemic domains, we would soon realise that the degree of such analysis is limited in virtue of the aptness of performances being epistemic in nature, i.e., there is a kind of epistemic perspective needed to discern whether the seat, shape, and situation of one's competence are arranged in such a way to facilitate aptness. This does not mean that epistemic domains are uniform in nature, as standards between epistemic domains differ greatly depending on the setting (E. Sosa 2021, 14).

When an LLM responds to a human input it is attempting to do something. The LLMs performance aims to predict a series of words that provide the human with a legible answer that appropriate to the prompt it receives.[174] The LLMs level of competence in the relevant domain impacts the likelihood of the LLM successfully making their attempt, and in cases where it attempts to aim at making a prediction that is accurate (true) according to the training data provided, its performance falls within an epistemic domain.[175]

If AI responses can manifest the AI's competence to some degree, then it seems fruitful to examine AI competence further using the SSS-model. Recall that SSS stands for seat, shape, and situation. In the case of a typical narrow AI, such as chess engines, we find that the seat, or constitutional competence, is not only in place, but oftentimes exceeds the highest performing

[174] ChatGPT, when prompted, replies that it adheres to guidelines to ensure that its responses are helpful, respectful, and safe. These guidelines are then further broken down into accuracy, relevance, respectfulness, safety, privacy, neutrality, transparency, ethical considerations, legal compliance, and empathy (OpenAI, ChatGPT 2024).

[175] Note here that, strictly speaking, the performance of LLMs like ChatGPT is not intended to inform, but to predict the next word in a string of text. However, it is not far-fetched to claim that the predictions are made to correspond to the guidelines which include parameters, one of which is accuracy.

humans in the relevant domain. What about NAI systems that are able to reach across domains? LLMs are capable of writing poetry and speeches, it is fluent in nearly every programming language, and it has access to immense amounts of data which it can quickly sort through to answer questions. One way to answer whether LLMs have constitutional competence is by examining whether the LLMs would be able to perform reliably enough[176] if they were in proper shape and in an appropriate situation (E. Sosa 2010, 473).[177]

So, let us examine LLMs shape and situation and see how applicable they are to artificial intelligence before we make a judgement about LLM constitutional competence. The shape, or inner competence, of NAI in general, including LLMs, seems unlike the inner competence humans possess because many of the examples generally used to describe this kind of competence, such as sobriety and alertness, do not apply to NAI. We can still imagine that a proper shape for an AI would consist of being configured correctly. This means the AI would not only have the appropriate nodes on the neural network after the relevant machine learning processes, but also the kind of configuration that enables it to correctly determine the appropriate response. To illustrate, think of an author that, when sober, can distinguish between fact and fiction, but when drunk they confuse the two and claim something as true that only happened in their book. In much the same way, LLMs have access to a vast amount of data, some of which is factual and some of which is fictional.

Another way in which AI's shape could be compromised is when it gets stuck in a local minimum, unable to successfully perform. For clarification, think about a chess engine trying to find the best move. It sees the potential moves on the board and immediately dismisses the moves that do not seem promising. One such move is sacrificing the queen for apparently no compensation. As it turns out, that queen sacrifice is the best move at a very high depth. Unfortunately, the NAI is stuck with the candidate moves it initially decided on, trying to figure out which of them is most promising. In virtue of the algorithms and the neural network the LLM contains, the LLM can access, transform, and provide whatever data is needed. However,

[176] "Reliably enough" is determined by the norms of the relevant domains of performance, for a detailed discussion on this, see Sosa (2010).
[177] This framing is inspired by Carter's (2018) approach in *Virtue Epistemology, Enhancement and Control*.

without a protocol to clearly distinguish between the two the LLM is not in proper shape to successfully perform in the epistemic domain of transmitting truths.

Regarding the appropriate situation, we find that LLMs rely heavily on the data they are fed with, and corrupted data can lead to false responses. If the situation is appropriate, i.e., the data is good, and the LLM is properly configurated, LLMs are clearly capable of performing well, with the caveat that their performance must be within their domain of expertise. In sum, NAI such as LLMs can possess complete competence if we accept that consciousness is not a necessary factor of proper shape and is rather determined by whether an epistemic agent is able to access the skills they harbour.

The implication here is that NAI can at least portray something that corresponds to sub-credal animal knowledge, or apt alethic affirmations, in the sense that they predict that something is true, their prediction manifests their competence, and their prediction is true because of said competence. What about reflective knowledge? A requirement of reflective knowledge is that the aptness of the NAI's prediction would itself manifest a second-order competence. This kind of meta-competence is only present if it entails a second-order competence of judging whether the performance would be apt. Is NAI capable of such reflection? I argue that it cannot.

One of the problems that NAI, such as LLMs, face is that they can generate hallucinations while being unaware they are doing so, which is a strike against reflective knowledge. It suggests that they have an unreliable grasp on their own first-order competence. Furthermore, regardless of what data you add to the NAI system, it cannot prove that data to be correct, i.e., the data it is working with is not safe, and the system is incapable of reflecting on that safety. Sosa gives an example of Norm, the normal perceiver, and Abnor, the mental ward patient (E. Sosa 2021, 148). On any given day when Abnor wakes up, he could have been experimented on which would deprive him of any or all epistemic competences that day. Recall that these competences are split into non-global competences that host experiences or seemings, and global epistemic competence, which amounts to a holistic basic belief-forming competence required to properly make any judgement or forming any belief.[178] Now, when Abnor wakes up on a given day he

[178] In Sosa's view, whether such global competences exist is an empirical question (E. Sosa 2021, 148).

might be disabled in the sense that some or all of his non-global competences have been taken away, which is potentially undetectable. In this case we find it doubtful that Abnor really knows when he could have so easily been unknowingly mistaken. If that is not convincing enough, Abnor could also have been robbed of his global epistemic competence, in which case *there is no way for him to properly reflect on any tell-tale signs that this has happened* (E. Sosa 2021, 149).

According to Sosa, Norm is positioned to have both animal and reflective perceptual knowledge, while Abnor is only able to have animal knowledge (E. Sosa 2021, 151). If Abnor only loses his non-global competence, he is only able to acquire reflective knowledge if someone tells him that his non-global competence, be it shape or situation, has been disabled. If he loses his global epistemic competence, he would lack the ability to assess its presence (E. Sosa 2021, 151).[179] What Norm and Abnor can teach us here, is that NAI could potentially have something close to reflective knowledge if the AI architects would tip the NAI off by confirming that the apt prediction was in fact not aptly made, but it is difficult to accept that this would amount to reflective knowledge for two reasons.

First, even if the NAI would be tipped off about its unfavourable shape or situation, it cannot properly make use of that information to arrive at an apt prediction in the way Norm is able to. Second, it seems like the system architects are the ones reflecting in this case, which goes against the spirit of what reflective knowledge is meant to represent —viz., this kind of secondary-reflection does not make it less likely that the NAI would have predicted inaptly, and the NAI cannot be confident that a reflection of this sort is generally reliable considering it has to rely completely on the architects to intervene whenever it makes a prediction.

When it comes to the loss of global competence, we see that NAI is in the same or worse position as Abnor. At best it can only tell when its global epistemic competence has not been compromised with no way of knowing when it has, at worst it cannot even tell when it has not been compromised because it is not safe (E. Sosa 2021, 152). This suggests that NAI is vulnerable to environmental epistemic luck fake barn cases. Imagine that it had complete

[179] Although Sosa points out that if it is not missing then it is possible to know that it is not missing in virtue of having the necessary competence to know so (E. Sosa 2021, 152).

competence while making a prediction while being unable to verify the appropriateness of the configuration (shape) and data (situation). In that case the NAI would be susceptible to epistemic luck, where either the shape or situation would be epistemically unfavourable, but the prediction would still be true on the basis of environmental luck.

What kind of an AI would be able to reflect on its own competence? If an AI could assess whether its non-global competence is compromised, thus capable of reflecting on its own reliability, it could be classified as a general artificial intelligence, with human-like cognition in epistemic domains. Although NAI can perform well within its domain, it does not have complete competence in the epistemic domains. I propose the following definitions of NAI and GAI from competence:

> **Narrow artificial intelligence**: An AI that has complete competence in some domain other than epistemic domains.
>
> **General artificial intelligence**: An AI that has the constitutional competence in epistemic domains to reflect on its complete first-order competence.

Conclusion

Having a clear virtue-theoretic distinction between NAI and GAI enables us to assess AI systems using an established and robust virtue epistemological framework. We can see how NAI is limited to animal knowledge (non-global competence) regardless of its level of competence it maintains in a specific domain while GAI would potentially be able to possess the epistemic competence needed for reflective knowledge. Note that actual reflection is not a requirement of GAI. Consider that someone that is placed in a fake barn case possesses human-like cognition even when they are not able to aptly reflect on their apt belief. To have the innermost constitutional skills required to reflect is sufficient to claim a level of cognition generally reserved for humans, and this would be enough for an AI to be considered GAI.

Furthermore, we see that competence in a domain other than the epistemic is not a requirement of GAI, because knowing as an epistemic endeavour does not depend on competences in other

domains. This suggests that GAI is not simply an NAI system that has been improved by some degree, but rather a different kind of AI. That is, NAI cannot become a GAI by advancing their competence in their respective domains unless they advance in the epistemic domains. A surprising result here is that the development of GAI competence runs into many of the same problems epistemologists have been working on, such as the problem of environmental epistemic luck. If GAI's second-order competence is to be complete, it must be properly situated both in its prediction and in its reflection (E. Sosa 2010, 473). If this is correct, then NAI can obtain sub-credal knowledge while GAI can reach the lowest creditability for knowledge. Considering that although predictions are at its core affirmations that involve a varying degree of guesswork, the predictions made by GAI are aptly apt in a way that avoids the limitations of NAI, i.e., GAI can predict in a manner that constitutes reflective knowledge (global competence).

Having a theory of what constitutes general AI can be beneficial beyond simply serving as a threshold that marks a distinction between NAI and GAI. For example, there is a tight connection between intelligence and action. Some actions we consider accidental, like a boulder that hits another boulder after breaking off a cliff. Other actions have intention behind them, like raising one's arm. Both kinds of actions have cause and effect relationships, but only the latter can be said to be purposeful. One way to explain what makes an action purposeful as opposed to accidental in the involvement of intelligence, i.e., that an action is purposeful when it is guided by intelligence. For example, the difference between someone that trips accidentally and someone that trips on purpose, where both individuals fall identically, is that the latter action manifests intent because it is guided by intelligence. Having a theory of what constitutes general intelligence AI, which draws on Sosa's structure of reflective knowledge, can provide us with a principled way to assess AI outputs by distinguishing between purposeful AI behaviour and mere accidental but predictable causal patterns.

On a final note, we can see how this distinction between NAI and GAI can help us navigate the notion of trusting AI. The *Ethics Guidelines for Trustworthy Artificial Intelligence* of the European Commission's High-Level Expert Group on AI (HLEG) finds that "it is important to build AI systems that are worthy of trust" (AI HLEG 2019, 35). HLEG further finds that for a system to be trustworthy we must be able to understand its actions.

This implicates the research field of XAI, or explainable AI, which aims to understand the underlying mechanisms of AI (AI HLEG 2019, 21). Simion and Willard-Kyle (Forthcoming) suggest that it is misguided to think of explainable AI (XAI) as the key to rational trust in AI because it is not generally true that rational trust requires understanding why, instead opting for what they call *the simple view of AI trust*, in which rational trust requires not AI explainability, but AI trustworthiness (Simion and Willard-Kyle Forthcoming). Simion and Kelp (2023) have put forth a functionalist account of AI trustworthiness, where they argue that that AI is trustworthy when it fulfils its function. However, Carter (2023) points out that AI functions are narrow and domain-specific and as such might not entail the sort of general trustworthiness that Simion and Kelp suggest (2023, 6). Zanotti et al. (2024) maintain that trust between humans and AI shares a *conceptual core* with trust between humans, which motivates using a notion of trust rather than reliability in applications to AI systems.[180] Others, like Ryan (2020) and Al (2023), have called for abandoning the idea of trustworthy AI and instead approach it in terms of reliability, as a rational account of reliability does not rely on AI to have emotion towards the trustor (affective) or be responsible for its actions (normative) (Ryan 2020, 17).

Having a virtue-theoretic distinction between NAI and GAI as presented in this chapter can potentially contribute to this ongoing discussion. For one, it facilitates a division between the notion of trusting NAI on one hand, and GAI on the other. GAI being capable of attaining reflective knowledge suggests that they can be held responsible for their apt judgements, and thus have the potential to be trustworthy. This further entails that GAI can be blamed for not living up to our trust and that it would be appropriate to consider failures of GAI to execute their design function as betrayals rather than disappointments. For clarification, consider that as GAI would have competence in epistemic domains to reflect on their first-order competences, one of their design functions will be to attain the status of apt judgements. In short, when we trust GAI, we are not only trusting their apt predictions, but we are trusting that they are performing their function to assess whether they were aptly apt in their predictions. Conversely, the narrow and domain-specific functions of NAI restrict its potential as a

[180] Although they suggest a distinction between trust in AI systems and interpersonal trust (Zanotti, et al. 2024, 2691).

trustworthy artefact. However, NAI predictions can still be highly reliable due to its extraordinarily high ceiling of competence within their narrow set of domains.

Conclusion

The first chapter demonstrated that it is possible to transmit knowledge without shared intentions and showed that Greco's account of knowledge transmission, which depends on these shared intentions through joint agency, is challenged. I introduced counterexamples, involving unwanted knowledge transmissions, that present a dilemma for anti-reductionist knowledge transmission views that rely on shared intentions and joint agency for credit. Either they concede that joint agency is not a necessary condition for knowledge transmissions, or they insist that it is present in all cases of knowledge transmissions, including cases of unwanted knowledge transmission.

Accepting the first horn of the dilemma by conceding that unwanted knowledge transmissions are not knowledge transmissions because they lack shared intentions and thus joint agency, would exclude paradigmatic examples of knowledge transmissions, which in turn makes the transmission theories that rely on joint agency only applicable to a narrow subset of transmission cases. These cases of knowledge transmission would thus encounter Lackey's dilemma, as they do not exhibit competent joint agency that can be credited for the true belief of the hearer as a result of testimony.

Going for the second horn, claiming that unwanted knowledge transmissions are cases of knowledge transmission, means having to accept a diminished form of joint agency that would be a by-product present in all testimonial exchanges. This form of joint agency would be trivial, making it difficult to recognise what exactly makes knowledge transmission special. Further, accounts that accept joint agency in this diminished form would overgeneralise to cases of knowledge transmissions that are clearly not cooperative, and that do not follow Greco's shared intention principle. A weak notion of joint agency cannot explain how a success can be credited to the hearer's competent agency even if that agency is not the most salient part of the success, as the success is not directly attributable to this trivial version of joint agency.

The conclusion of the first chapter is that competent joint agency is either not present in many typical cases of testimonial knowledge exchanges, which makes those cases susceptible to

Lackey's dilemma, or it is present in a diminished form that cannot produce the sort of shared credit that is needed to defend against it.

In the second chapter, I argued for the basic credit view. It presents a novel way of defending the idea that credit always follows knowledge against counterexamples. I began by demonstrating how an individual-focused credit view is vulnerable to testimonial cases like CHICAGO VISITOR. I then provided definitions for testimonial knowledge chains and epistemic sources to motivate a social species of a credit view, inspired by speaker's accounts of knowledge transmission. The social credit view emphasises the role of the speaker in some testimonial cases. If correct, I could argue that speakers can deserve credit for the hearer's belief, as the speaker's cognitive abilities are a salient enough part of why the hearer came to know.

Although an initially promising view, I showed how a certain kind of counterexamples, that involve hearers that come to know after being told without the speaker possessing knowledge about what they testified, oppose such views.

Finally, I presented the basic credit view. I argued that there are basic epistemic achievements that might not be as valuable as knowledge or understanding but are nevertheless genuine achievements that are creditable as such. I further argued that having cognitive contact with reality from ability is one such achievement, as it is a cognitive success from ability that follows an achievement structure. On the basic credit view, testimonial knowledge requires there to have been cognitive contact with reality through ability somewhere in the testimonial chain. I argued that this allows us to defend that there is no knowledge without credit. When there is testimonial knowledge there is credit because someone in the testimonial knowledge chain exhibited cognitive contact with reality, a creditable achievement. With regards to non-testimonial cases, I argued that we do not have to restrict ourselves to the basic credit view, which only claims that credit follows knowledge. Instead, we can depend on the stronger claim of the individual credit view, that one knows if one has a true belief because of their cognitive abilities, because the challenging testimonial-based counterexamples are absent in such non-testimonial cases.

I acknowledged that the more demanding forms of epistemic achievements, such as knowledge and understanding, require more than cognitive contact with reality, for example a true belief from cognitive ability, and that the credit for such achievements is greater than the basic credit that can be attributed to basic epistemic achievements such as cognitive contact with reality from ability.

Furthermore, I identified similarities between paradigmatic counterexamples to transmission theories more generally, and how they are structured in a way that exploits two kinds of unreliability of epistemic agents. On one hand they can be unreliable in their testimonial exchanges, for example, by lying in the case of a speaker, and by mishearing what was testified in the case of a hearer. On the other they can be unreliable believers, making it possible for speakers to transmit unsafe true beliefs, and hearers to acquire true beliefs that are unsafe. Having identified these features, and how counterexamples can be constructed involving one or more of those unreliable properties of epistemic agents, I generated the different possible kinds of counterexamples and responded to them.

The basic credit view shows that there is no knowledge without credit, as the individual credit view, which the basic credit view holds on to in non-testimonial cases, provides an explanation for why knowledge is incompatible with intervening epistemic luck in non-testimonial cases. In the case of testimonial knowledge, someone in the testimonial chain exhibited some cognitive success from ability, which is a creditable basic epistemic achievement. The conclusion is that counterexamples like GRANT SHOLARS and FOSSIL do not pose a challenge to the fundamental claim of the extended family of credit views that knowledge always entails credit.

In the third chapter, I identified and introduced various temporal elements of trust and trust-related phenomena that are a product of the relationship between trust and time.

I presented a distinction between trust affirmation and ongoing trust. Trust affirmation can be characterised by a trustor's initial belief towards the trustee's trustworthiness, doxastically formed using evidential norms. I showed how precautionary measures of trust can be freely executed before trust affirmations without undermining the trust that follows. I further argued

that during trust affirmations there is no time to engage in precautionary measures such as monitoring, searching for evidence, or rational reflection.

I argued that trust affirmations are followed by ongoing trust, which can be characterised as trust that continues over a period of time. I then demonstrated that not all cases of ongoing trust react to precautionary action in the same manner. To explain why that is, I distinguished between trust resolutions and trust confirmations, where trust resolution occurs only if the trustee does as entrusted, and trust is confirmed when the trustor comes to know that the trust was resolved. Having made that distinction I then presented a distinction between definite trust and indefinite trust.

I showed that definite trust involves cases where trust has a terminus, such as trusting someone to perform a specific task within a set timeframe. I found that therapeutic trust cases must necessarily be instances of definite trust because of what I call confirmation monitoring, which can be characterised as a kind of monitoring that results in trust confirmations. Therapeutic trust cases can be successful as long as they aim to build trust, and we argued that confirmation monitoring is necessary for any therapeutic increases in trust. Following this, I argued that successful therapeutic trust cases are always cases of definite trust.

I then explored how precautionary actions and temporal elements of trust relate. I analysed the risk-mitigating effects of precautionary actions and distinguished between the risk of betrayal and the risk of disutility in cases where the entrusted action was not performed. I argued that precautionary measures can partially guard against the risk of the entrusted action not being completed, but that it cannot guard against the risk of being betrayed. I concluded that one cannot fully guard against the risk without eliminating the trust, because precautionary actions that would fully protect against the risk of disutility would indicate a complete lack of trust and are thus incompatible with trust in a way that weaker precautionary actions are not.

I then challenged prominent trust accounts using the distinctions made in the chapter. I demonstrated that monitoring does not always undermine trust, a controversial claim that challenges Keren's doxastic trust account, as it claims that reasons for trust are second-order preemptive reasons against engaging in precautionary actions like monitoring or reflection.

Furthermore, I argued that preemptive reasons are not always suited to minimise risk, often having a minimal effect on the quality of the trusting relation.

I concluded that preemptive reasons are not present during trust affirmations, but essential in ongoing indefinite trust cases, with some caveats. I further concluded that preemptive reasons must necessarily be ignored for successful trusting in ongoing definite trust cases, including cases that are generally referred to as therapeutic trust cases.

In the following section, I examined Emma Gordon's claim that monitoring can facilitate trust. Although I agreed with Gordon's general assessment that monitoring can be beneficial to trust, and thereby that there is less tension between trust and monitoring than commonly thought, we came to the same conclusion for different reasons. Using the distinctions I made, I argued that when trust is rebuilt after being broken, it must be reaffirmed, thus causing a new trusting relation to begin. One consequence of this is that it explains why monitoring seems compatible with trust in the way Gordon claims, namely, because in between the terminus of the previous ongoing trust and the trust reaffirmation is a period without trust. During this period, epistemic agents are free to monitor because there is no trust to undermine. I still argued that the overarching claim is correct, that monitoring can facilitate trust, because confirmation monitoring is a necessary feature of successful therapeutic trust increases and definite trust cases more generally.

In the last section of the third chapter, I argued against Wanderer and Townsend's attempt to reconcile trust and rationality. They accepted the tension between trust and rationality by claiming that the tension is genuine and ineliminable, and that rational doubts can make trust all the more commendable. I agreed that indefinite trust is undermined by monitoring in the way Wanderer and Townsend suggest. However, I demonstrated that the tension in definite trust cases is limited due to confirmation monitoring. This monitoring, whether accidental, which I defined earlier in the chapter as pseudomonitoring, or deliberate, plays an important role in the success of definite trust cases. I then argued that vexed trust, as presented by Wanderer and Townsend, is less desirable than unvexed trust. The reason is that vexed trust results from doubt, and I argued that reflecting on trust undermines it, unlike confirmation monitoring. Finally, I showed how, when comparing vexed and unvexed trustors, in cases

where two individuals trust, one vexed and the other unvexed, less is needed for the vexed truster to stop trusting, further indicating that unvexed trust is less fragile.

In the fourth chapter, I argued that groups that employ gatekeeping by relying on distribution norms in testimonial exchanges within the group to facilitate cheap knowledge transmissions, are at risk of incurring two distinct problems. First, that groups can become "poisoned" when a member of the group forms a false belief and then proceeds to share the misinformation within the group, thinking that it is good information. The other members of the group are placed in an epistemically vulnerable position, as they will be less critical of the testimony they hear from one of their own. This means that a single false belief can easily permeate throughout a gatekeeping group. If the same happens to a group that does not employ gatekeeping, that is, does not use distribution norms within the group, then the false belief would not move between members so easily.

The second problem is that a group member might dismiss good information, believing it to be bad. I argue that this can lead to knowledge deprivation as a result of insufficient coverage, as group members rely on each other to filter good information from bad, and when a member dismisses good information, the entire group is deprived of it.

I argue that groups face a challenge when they attempt to counter these two problems of gatekeeping. In order for groups to lower the risks of their gatekeeping failing, they must increase the epistemic labour of their gatekeeping by shifting their norms further to the acquisition side. This would make knowledge transmissions within the group more expensive, understood as an increase in the epistemic work required to share information, resulting in gatekeeping that is safer but less efficient. The only way to eliminate the risk of the two problems is to rely solely on acquisition norms, neutralising gatekeeping completely. With this in mind, it seems difficult to justify these epistemic harms for cheaper transmissions, and by extension gatekeeping.

I then argue that if a certain set of conditions are met, gatekeeping can be justified. I proposed three conditions for justified gatekeeping that aim to minimise the risks involved by focusing on reducing the impact of the problems if they occur, rather than minimising the probability of

the risks realising. The conditions for justified epistemic gatekeeping are: (1) the group must be a part of a collective of groups that form an epistemically interrelated community, (2) the group is sufficiently epistemically autonomous, and (3) the groups within the community must be at least subjectively epistemically rational, that is, they have the goal of believing truths, avoiding falsehoods, and abide by epistemic practices that they believe are effective in achieving that goal.

Together, these conditions aim to ensure that the community of groups epistemically prospers without incurring long-term epistemic harms to individual groups within it. I concede that these conditions cannot eliminate the risk gatekeeping groups face, but I maintain that communities of groups are better protected under these conditions than they would be without them.

I conclude the chapter by demonstrate that the three conditions are most naturally fulfilled by expert communities, such as the scientific community, because of how they are structured and the shared goals of the groups within such communities. These shared truth-conducive goals encourage them to correct their course in case of gatekeeping failures.

In the fifth and final chapter, I presented a virtue-theoretic approach to artificial intelligence (AI) systems by identifying parallels between human competence and AI competence. I then illustrated how this framework can be used to establish what separates general AI (GAI) from narrow AI (NAI) from a virtue-theoretic perspective. I argue that a base distinction can be made between GAI and NAI by examining the range of domains the AI is proficient in. While NAI is a task-specific optimised system in its relevant domain, GAI would portray general cognitive autonomy in a variety of domains.

Before moving on to a virtue-theoretic distinction, I addressed worries relating to the fact that NAI only indirectly and incidentally tracks truths, which could be problematic when arguing for AI competence by comparing AI attempts at performance with human attempts. I responded by pointing out that epistemic competences have non-epistemic function-generated aims, and that this does not stand in tension with the idea that our epistemic competences reliably deliver accurate information. If a large language model (LLM), a type of NAI, is optimised to produce plausible-sounding content, and that is its design function, then that does not preclude it from

having the kind of reliable connection with truth that epistemic competence demands. I then addressed a different worry, that AI is thought to be incapable of possessing beliefs. I responded by pointing out that Herman Cappelen and Josh Dever (2021) have developed a de-anthropocentrised externalism that suggests that AI can make statements that predicate properties. Furthermore, I argued that as long as we assume that the relevant AI is at least configured and optimised to make accurate predictions, then we can claim that AI can possess competence. I then made a comparison between guesses and LLM predictions, as guesses, when they manifest competence within a virtue-theoretic assessment, can be apt.

I referred to a virtue-theoretic distinction between two kinds of epistemic competences, namely, non-global competences which host seemings, and global epistemic competences, which are judgment-forming competences. Then, I argued that, because global epistemic competences are used in determining how to judge "all things considered", that for an AI to possess this sort of holistic global epistemic competence it would need to be capable of making judgments "all things considered". I further argued that NAI can possess the kind of non-global competence that corresponds to sub-credal animal knowledge, or apt alethic affirmations, in the sense that they predict that something is true, their prediction manifests their competence, and their prediction is true because of said competence. Then, I demonstrated that NAI is not capable of making apt predictions that would themselves manifest a second-order competence.

I argued that for an AI to be considered a GAI, it must have the constitutional competence in epistemic domains to reflect on its complete first-order competence. Meanwhile, NAI can be defined as having complete competence in some non-epistemic domain. This suggests that GAI is not simply an NAI system that has been improved by some degree, but rather a different kind of AI. That is, NAI cannot become a GAI by advancing their competence in their respective domains unless they advance in the epistemic domains. Furthermore, I argued that GAI could reach the lowest creditability for knowledge because the predictions made by GAI are aptly apt in a way that avoids the limitations of NAI, i.e., GAI can predict in a manner that constitutes reflective knowledge (global competence).

Further exploration

An area for further exploration concerns both definite and indefinite trust. One question that has not been addressed fully in this thesis is whether indefinite trust instances can be resolved. Arguably not, if we agree with the definitions and ideas presented in this thesis, which state that the capacity to be resolved is a distinguishing feature of definite trust. The problem is that instances of indefinite trust seem to possess unique properties that can have implications for how we view both indefinite and definite trust. Consider the following questions:

1. Can indefinite trust become definite?
2. Does human mortality affect trust?

On the first question, in chapter three I examined whether it would be possible to conceive of indefinite trust as a series of definite trust instances, but I did not fully address what happens when indefinite trust instances end. By definition indefinite trust instances cannot be resolved, but they can still end. Consider a couple who trusts each other to be faithful (and they have been faithful) but breaks up for any reason other than infidelity. Their trust is a paradigmatic example of indefinite trust and cannot be resolved. Yet, when they break up, they have in some odd way resolved the trust; they did not cheat on each other during the relationship, and they fulfilled their end of the bargain. The trust disappeared when they broke up, as they could not betray that trust after separating even if they wanted to, but that is no different from definite trust cases that have been confirmed; I could not betray your trust by not cleaning the dishes as I promised if I have already cleaned the dishes.

If we consider that the couple's trust was really a kind of definite trust case, because they trusted each other to be faithful up until the relationship ended. However, this would suggest that the relationship had a specific end point, and if it does not then it seems more accurate to think of their trust as trusting each other to be faithful while they are together. This discussion relates to our second question.

The second question concerns instances when human mortality becomes a consideration of trust in a unique way. For example, if I trust you to do something at a certain time, which would be

a typical case of definite trust, but you passed away before you could perform the entrusted action. Did you betray my trust? The obvious response is no, of course not. However, you did not do as entrusted, even though you might have had the right kind of intentions. It is not unusual to forgive someone after they have passed away. Perhaps this is one of those instances in which I was betrayed, but because I cannot blame you for why you did not do as entrusted, it is easy to forgive (so easy in fact that we do not consider it betrayal). This raises further questions, for example, how can it be that if I trust you to do something before a certain time, and then realise you have not done so, I feel betrayed immediately, but when I find out you were in a minor accident and had to go to the hospital, I stop feeling betrayed. Strangely, I might then inquire about when exactly the accident happened and whether you could have done as entrusted before that time. Similarly, I might want to know what you were doing that caused the accident before deciding whether I should feel betrayed.

A different example that relates to the second question is someone on their death bed trusting a close friend to make certain funeral arrangements. This kind of trust affirmation can only become ongoing trust for a brief moment, and the trustor has no way to monitor the trustee or follow through with their part of the trusting relation, i.e., confirming that the trust was resolved. Does this indicate that people close to dying are unable to trust others because they cannot take precautionary measures (even if they wanted to), they cannot be harmed by trust failures (as the realised risk will not affect them posthumously), and they cannot be betrayed in the traditional sense? One line of response would be to simply ascribe hope in lieu of trust to cases like these. Another response would be to point out that people often say things like “their legacy was betrayed”, or that “their dying wish was not honoured”, which at least indicates that there is some posthumous risk involved.

Bibliography

Adler, Jonathan E. 1994. “Testimony, trust, knowing.” *Journal of Philosophy* 91 (5): 264-275. doi:10.2307/2940754.

AI HLEG. 2019. “Ethics guidelines for trustworthy AI.” *European Commission.* 8 April. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Al, Pepijn. 2023. “(E)-Trust and Its Function: Why We Shouldn’t Apply Trust and Trustworthiness to Human-AI Relations.” *Journal of Applied Philosophy* 40 (1): 95-108. doi:doi: 10.1111/japp.12613.

Audi, Robert. 2006. “Testimony, Credulity, and Veracity.” In *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, 25-49. Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199276011.003.0002.

Audi, Robert. 1997. “The Place of Testimony in the Fabric of Knowledge and Justification.” *American Philosophical Quarterly* 34: 405–422. https://www.jstor.org/stable/20009910.

Baier, Annette. 1986. “Trust and Antitrust.” *Ethics* 96 (2): 231-260. doi:https://doi.org/10.1086/292745.

Baker, Judith. 1987. “Trust and Rationality.” *Pacific Philosophical Quarterly* 68 (1): 1-13. doi: https://doi.org/10.1111/j.1468-0114.1987.tb00280.x.

Baker, Judith, and Philip Clark. 2018. “Epistemic buck-passing and the interpersonal view of testimony.” *Canadian Journal of Philosophy* 48 (2): 178-199. doi:https://doi.org/10.1080/00455091.2017.1341781.

Barandiaran, X. E., E. Di Paolo, and M. Rohde. 2009. “Defining Agency: Individuality, normativity, asymmetry, and spatio-temporality in action.” *Journal of Adaptive Behavior* 17 (5): 367-386. doi:10.1177/1059712309343819.

Barimah, George Kwasi. 2024. “Epistemic Trust in Scientific Experts: A Moral Dimension.” *Science and Engineering Ethics* 30 (21): 1-22. doi:https://doi.org/10.1007/s11948-024-00489-x.

Barzilai-Nahon, Karine. 2008. “Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control.” *Journal of the American Information Science and Technology* 59 (9): 1–20. doi:https://doi.org/10.1002/asi.20857.

BonJour, Laurence. 1980. “Externalist Theories of Empirical Knowledge.” *Midwest Studies in Philosophy* 5: 53–73. doi:https://doi.org/10.1111/j.1475-4975.1980.tb00396.x.

Booth, Anthony Robert. 2018. “Trust in the Guise of Belie.” *International Journal of Philosophical Studies* 1-17. doi:10.1080/09672559.2018.1450075.

Bratman, Michael E. 1992. “Shared Cooperative Activity.” *The Philosophical Review* (Duke University Press) 101 (2): 327-341. doi:https://doi.org/10.2307/2185537.

Broncano-Berrocal, Fernando, and J. Adam Carter. 2020. “Deliberation and Group Disagreement.” In *The Epistemology of Group Disagreement*, edited by Fernando Broncano-Berrocal and J. Adam Carter, 9-45. New York: Routledge. doi:https://doi.org/10.4324/9780429022500.

—. 2020. *The Epistemology of Group Disagreement.* Edited by Fernando Broncano-Berrocal and J. Adam Carter. New York: Routledge. doi:https://doi.org/10.4324/9780429022500-1.

Burge, Tyler. 1993. “Content Preservation.” *The Philosophical Review* 102: 457–88. doi:https://doi.org/10.2307/2185680.

Campbell, Keith. 1970. *Body and Mind.* Red Globe Press London. doi:https://doi.org/10.1007/978-1-349-00678-6.

Cappelen, Herman, and Josh Dever. 2021. *Making AI Intelligible: Philosophical Foundations.* New York: Oxford University Press. doi:https://doi.org/10.1093/oso/9780192894724.001.0001.

Carter, J. Adam. 2024. *A Telic Theory of Trust.* Oxford University Press. doi:https://doi.org/10.1093/9780191982460.001.0001.

Carter, J. Adam. 2020. “On Behalf of a Bi-Level Account of Trust.” *Philosophical Studies* (177): 2299–2322. doi:https://doi.org/10.1007/s11098-019-01311-2.

Carter, J. Adam. 2023. “Simion and Kelp on trustworthy AI.” *Asian Journal of Philosophy* 2 (18): 1-8. doi:https://doi.org/10.1007/s44204-023-00067-1.

—. 2024. *Stratified Virtue Epistemology: A Defence.* Cambridge University Press: Cambridge.

Carter, J. Adam. 2022. “Therapeutic trust.” *Philosophical Psychology* 37 (1): 38-61. doi:10.1080/09515089.2022.2058925.

Carter, J. Adam. 2018. “Virtue Epistemology, Enhancement and Control.” *Metaphilosophy* 283-304. doi:https://doi.org/10.1111/meta.12304.

Carter, J. Adam, and Clayton Littlejohn. 2021. *This is Epistemology: An Introduction.* Edited by Clayton Littlejohn. Hoboken: Wiley-Blackwell.

Carter, J. Adam, and Philip J. Nickel. 2014. “On Testimony and Transmission.” *Episteme* 11 (2): 145-155. doi:doi:10.1017/epi.2014.4.

Cassinadri, Guido. 2024. “ChatGPT and the Technology-Education Tension: Applying Contextual Virtue Epistemology to a Cognitive Artifact.” *Philosophy & Technology* 37 (14): 1-28. doi:https://doi.org/10.1007/s13347-024-00701-7.

Castelfranchi, Christiano, and Rino Falcone. 2000. “Trust and control: A dialectic link.” *Applied Artificial Intelligence* 14 (8): 799-823. doi:10.1080/08839510050127560.

Chalmers, David. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.

Coady, C. A. J. 1992. *Testimony: A Philosophical Study.* Oxford: Clarendon Press.

Craig, Edward. 1990. *Knowledge and the State of Nature.* Oxford: Oxford University Press.

Dellsén, Finnur. 2021. “Consensus versus Unanimity: Which Carries More Weight?” *British Journal for the Philosophy of Science* 1-30. doi:https://doi.org/10.1086/718273.

Dellsén, Finnur. 2020. “The epistemic value of expert autonomy.” *Philosophy and Phenomenological Research* 100 (2): 344-361. doi:https://doi.org/10.1111/phpr.12550.

Dennett, Daniel. 1991. *Consciousness Explained.* The Penguin Press.

Dormandy, Katherine. 2019. “Introduction: An Overview of Trust and Some Key Epistemological Applications.” In *Trust in Epistemology,* edited by Katherine Dormandy, 1-40. New York: Taylor & Francis. doi:https://doi.org/10.4324/9781351264884-1.

Dormandy, Katherine, and Bruce Grimley. 2024. “Gatekeeping in Science: Lessons from the Case of Psychology and Neuro-Linguistic Programming.” *Social Epistemology* 38 (3): 392-412. doi:10.1080/02691728.2024.2326828.

Dougherty, Tom. 2013. “Sex, Lies, and Consent.” *Ethics* 123 (4): 717-744. doi:https://doi.org/10.1086/670249.

—. 2021. *The Scope of Consent.* Oxford: Oxford University Press. doi:https://doi.org/10.1093/oso/9780192894793.001.0001.

Downie, R.S. 1963. “Hope.” *Philosophy and Phenomenological Research* 24 (2): 248-251. doi:https://doi.org/10.2307/2104466.

Dretske, Fred I. 1982. “A Cognitive Cul-de-sac.” *Mind* 91 (361): 109–111.

Eder, Anna-Maria Asunta. 2020. “Disagreement in a Group: Aggregation, Respect for Evidence, and Synergy.” In *The Epistemology of Group Disagreement*, edited by Fernando Broncano-Berrocal and J. Adam Carter, 184-210. New York: Routledge. doi:https://doi.org/10.4324/9780429022500-10.

Edwards, Benj. 2023. *Why ChatGPT and bing chat are so good at making things up.* 6 April. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/.

Elgin, Catherine Z. 2020. “Epistemic Gatekeepers.” In *The Aesthetics of Science*, by Catherine Z. Elgin, edited by M. Ivanova and S. French, 21-35. New York: Routledge. doi:https://doi.org/10.4324/9780429030284-2.

Faulkner, Paul. 2007. “A Geneology of Trust.” *Episteme* 4 (3): 305-321. doi:https://doi.org/10.3366/E174236000700010X.

—. 2011. *Knowledge on Trust.* Oxford: Oxford University Press.

Faulkner, Paul. 2007. “On Telling and Trusting.” *Mind* 116 (464): 875-902. doi:10.1093/mind/fzm875.

Faulkner, Paul. 2006. "Understanding knowledge transmission." *Ratio* 2 (19): 156-175. doi:https://doi.org/10.1111/j.1467-9329.2006.00317.x.

Floridi, Luciano, and Massimo Chiriatti. 2020. “GPT-3: its nature, scope, limits, and consequences.” *Mind Mach* 30 (4): 681-694. doi:https:// doi. org/ 10. 1007/s11023-020- 09548-1.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.” *Minds and Machines* 28: 689-707. doi:https://doi.org/10.1007/s11023-018-9482-5.

Foer, Franklin. 2017. *World Without Mind: The Existential Threat of Big Tech.* New York: Penguin Press.

Foley, Richard. 2001. *Intellectual Trust in Oneself and Others.* Cambridge University Press.

Fricker, Elizabeth. 2006. “Second-Hand Knowledge.” *Philosophy and Phenomenological Research* LXXIII (3): 592-618. doi:https://doi.org/10.1111/j.1933-1592.2006.tb00550.x.

Fricker, Elizabeth. 1987. “The Epistemology of Testimony.” *Proceedings of the Aristotelian Society* 61: 57-84.

Fried, Charles. 1981. *Contract as Promise.* Cambridge, MA: Harvard University Press.

Friedman, Ann. 2023. *The Real Meaning of Gatekeeping.* 31 July. Accessed August 2, 2024. https://www.thecut.com/2023/07/what-does-gatekeeping-mean.html.

Frost-Arnold, Karen. 2013. “Moral trust & scientific collaboration.” *Studies in History and Philosophy of Science Part A* 44 (3): 301–310. doi:https://doi.org/10.1016/j.shpsa.2013.04.002.

Frost-Arnold, Karen. 2014. “The cognitive attitude of rational trust.” *Synthese* 191: 1957–1974. doi:https://doi.org/10.1007/s11229-012-0151-6.

Gettier, Edmund L. 1963. "Is Justified True Belief Knowledge?" *Analysis, Vol. 23, No. 6* 121-123.

Gibbard, Alan. 1990. *Wise Choices, Apt Feelings.* Harvard University Press.

Gilbert, Margaret. 2009. “Shared intention and personal intentions.” *Philosophical Studies* 144 (1): 167–187. doi:https://doi.org/10.1007/s11098-009-9372-z.

Gilbert, Margaret. 1990. “Walking Together: A Paradigmatic Social Phenomenon.” *Midwest Studies in Philosophy* XV (1): 1-14. doi:https://doi.org/10.1111/j.1475-4975.1990.tb00202.x.

Ginet, Carl. 2001. “Deciding to believe.” In *Knowledge, Truth and Duty*, edited by Matthias Steup, 63-76. New York: Oxford University Press. doi:https://doi.org/10.1093/0195128923.003.0005.

Glass, Shirley P. 2007. *Not” Just Friends”: Rebuilding Trust and Recovering Your Sanity after Infidelity.* Simon and.

Goertzel, Ben, and Cassio Pennachin. 2007. “Contemporary Approaches to Artificial General Intelligence.” In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio

Pennachin, 1-28. Heidelberg: Springer Berlin. doi:https://doi.org/10.1007/978-3-540-68677-4.

Goldberg, Sanford C. 2012. “Epistemic extendedness, testimony, and the epistemology of instrument-based belief.” In *Foundations and Applications of Social Epistemology: Collected Essays*, by Sanford C. Goldberg, 213–234. Oxford: Oxford Academic. doi:https://doi.org/10.1093/oso/9780198856443.003.0013.

—. 2021. *Foundations and Applications of Social Epistemology.* Oxford: Oxford University Press. doi:https://doi.org/10.1093/oso/9780198856443.001.0001.

Goldberg, Sanford C. 2006. “Reductionism and the Distinctiveness of Testimonial Knowledge.” In *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, 127-144. Oxford: Oxford University Press.

Goldberg, Sanford C. 2005. “Testimonial knowledge through unsafe testimony.” *Analysis* 65: 302–311.

Goldberg, Sanford C. 2001. “Testimonially Based Knowledge from False Testimony.” *The Philosophical Quarterly* 51 (205): 512-526.

Goldberg, Sanford C. 2011. “The Division of Epistemic Labor.” *Episteme* (Edinburgh University Press) 8 (1): 112-125. doi:https://doi.org/10.3366/epi.2011.0010.

Goldman, Alvin I. 2014. “Social Process Reliabilism: Solving Justification Problems in Collective Epistemology.” In *Essays in Collective Epistemology*, edited by Jennifer Lackey, 11-41. Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199665792.003.0002.

Gordon, Emma C. 2022. “When Monitoring Facilitates Trust.” *Ethical Theory and Moral Practice* 25: 557-571. doi:https://doi.org/10.1007/s10677-022-10286-9.

Graham, Peter J. 2006. “Can Testimony Generate Knowledge?” *Philosophica* 78 (2): 105-127.

Graham, Peter J. 2000. “Conveying information.” *Synthese* 123: 365-392.

Graham, Peter J. 2016. “Testimonial Knowledge: A Unified Account.” *Philosophical Issues* 172-186. doi:doi: 10.1111/phis.12082.

Graham, Peter J. 2000. “Transferring knowledge.” *Noûs* 34 (1): 131–152.

Greco, John. 2012. “A (Different) Virtue Epistemology.” *Philosophy and Phenomenological Research* 85: 1-26.

—. 2010. *Achieving Knowledge.* Cambridge: Cambridge University Press.

Greco, John. 2004. "Knowledge as Credit for True Belief." In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, edited by Michael DePaul and Linda Zagzebski. Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199252732.003.0006.

Greco, John. 2012. “Recent Work on Testimonial Knowledge.” *American Philosophical Quarterly* (University of Illinois Press on behalf of the North American Philosophical Publications) 49 (1): 15-28. https://www.jstor.org/stable/23212646.

Greco, John. 2015. “Testimonial Knowledge and the Flow of Information.” In *Epistemic Evaluation: Purposeful Epistemology*, edited by Henderson David and Greco John, 274–290. Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199642632.003.0012.

—. 2020. *The Transmission of Knowledge.* Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/9781108560818.

Greco, John. 2009. “The Value Problem.” In *Epistemic Value*, by Adrian Haddock, Alan Millar and Duncan Pritchard, 313-322. Oxford: Oxford University Press.

Greco, John. 2008. “What's Wrong with Contextualism?” *The Philosophical Quarterly* 58 (232): 416-436. doi:https://doi.org/10.1111/j.1467-9213.2008.535.x.

Greco, John, and John Turri. 2011. *Virtue Epistemology.* Accessed February 14, 2017. https://plato.stanford.edu/entries/epistemology-virtue/.

Hardin, Russell. 1992. “The street-level epistemology of trust. 14 (2):.” *Analyse & Kritik* 14 (2): 152-176. doi:https://doi.org/10.1515/auk-1992-0204.

—. 2002. *Trust and trustworthiness.* New York: Russell Sage Foundation. doi:https://doi.org/10.1007/s11615-003-0046-8.

Hardwig, John. 1991. “The role of trust in knowledge.” *Journal of Philosophy* 88 (12): 693-708. doi:https://doi.org/10.1017/9781108560818.

Hawley, Katherine. 2014. “Trust, Distrust and Commitment.” *Noûs* 48 (1): 1-20. doi:https://doi.org/10.1111/nous.12000.

Helm, Bennett. 2008. “Plural Agents.” *Noûs* 42 (1): 17-49. doi:doi/10.1111/j.1468-0068.2007.00672.x.

Henderson, David. 2011. “Gate-Keeping Contextualism.” *Episteme* 8 (1): 83-98.

Hertlein, K.M., C. Dulley, R. Cloud, D. Leon, and Jenna Chang. 2017. “Does absence of evidence mean evidence of absence? Managing the issue of partner surveillance in infidelity treatment.” *Sexual and Relationship Therapy* 32 (3-4): 323-333. doi:10.1080/14681994.2017.1397952.

Herzog, Christian. 2022. “On the risk of confusing interpretability with explicability.” *AI and Ethics* 2: 219-225. doi:https://doi.org/10.1007/s43681-021-00121-9.

Hicks, M.T., J. Humphries, and J. Slater. 2024. “ChatGPT is bullshit.” *Ethics Inf Technol* 26 (38). doi:https://doi.org/10.1007/s10676-024-09775-5.

Hieronymi, Pamela. 2008. “The reasons of trust.” *Australasian Journal of Philosophy* 86 (2): 1-24. doi:https://doi.org/10.1080/00048400801886496.

Hills, Allison. 2009. “Moral Testimony and Moral Epistemology.” *Ethics* 120: 94–127.

Hinchman, Edward S. 2021. “Disappointed Yet Unbetrayed: A New Three-Place Analysis of Trust.” In *Social Trust,* edited by K. Vallier and M. Weber, 73-101. Routledge. doi:https://doi.org/10.4324/9781003029786-6.

Hinchman, Edward S. 2005. “Telling as Inviting to Trust.” *Philosophy and Phenomenological Research* (International Phenomenological Society) LXX (3): 562-587. https://www.jstor.org/stable/40040817.

Holton, Richard. 1994. “Deciding to trust, coming to believe.” *Australasian Journal of Philosophy* 72 (1): 63-76. doi:https://doi.org/10.1080/00048409412345881.

Horsburgh, H.J.N. 1960. “The Ethics of Trust.” *The Philosophical Quarterly* 10 (41): 343-354. doi:https://doi.org/10.2307/2216409.

Jespersen, Otto. 1924. *The Philosophy of Grammar.* London: George Allen & Unwin Ltd.

Jones, Karen. 2004. “Trust and Terror.” *Moral Psychology: Feminist Ethics and Social Theory* 3-18.

Jones, Karen. 1996. “Trust as an affective attitude.” *Ethics* 107 (1): 4-25. doi:https://doi.org/10.1086/233694.

Jones, Karen. 2012. “Trustworthiness.” *Ethics* 123 (1): 61-85. doi:https://doi.org/10.1086/667838.

Kallestrup, Jesper. 2022. “Nonreductive Group Knowledge Revisited.” *Episteme* 1-24.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, and Maria Bannert. 2023. “ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education.” *Learning and Individual Differences* 103. doi:https://doi.org/10.1016/j.lindif.2023.102274.

Kelp, Christoph, and Mona Simion. 2022. *Sharing Knowledge: A Functionalist Account of Assertion.* Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/9781009036818.

Kenny, Anthony. 1963. *Action, Emotion, and Will.* London: Routledge.

Keren, Arnon. 2014. “Trust and Belief: A Preemptive Reasons Account.” *Synthese* 191 (12): 2593-2615. doi:http://dx.doi.org/10.1007/s11229-014-0416-3.

Keren, Arnon. 2019. “Trust, Preemption, and Knowledge.” In *Trust in Epistemology*, 114-135. New York: Taylor & Francis. doi:https://doi.org/10.4324/9781351264884-5.

Kirk, Robert. 1974. “Sentience and Behaviour.” *Mind* 83: 43-60.

Kopec, Matthew. 2019. “Unifying Group Rationality.” *Ergo* 6 (18): 517-544. doi:https://doi.org/10.3998/ergo.12405314.0006.018.

Kvanvig, Jonathan. 2003. *The Value of Knowledge and the Pursuit of Understanding.* New York: Cambridge University Press.

Kwong, Jack M. C. 2023. “Gatekeeping the Mind.” *Inquiry: An Interdisciplinary Journal of Philosophy* 1-24. doi:https://doi.org/10.1080/0020174x.2022.2163284.

Lackey, Jennifer. 2006. “Knowing from Testimony.” *Philosophy Compass* (Blackwell Publishing) 1 (5): 432-448. doi:https://doi.org/10.1111/j.1747-9991.2006.00035.x.

Lackey, Jennifer. 2009. "Knowledge and credit." *Philosophy Studies* (142): 27-42.

—. 2008. *Learning from Words: Testimony as a Source of Knowledge.* Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199219162.001.0001.

Lackey, Jennifer. 2011. "Testimony: Acquiring Knowledge from Others." In *Social Epistemology: Essential Readings*, edited by Alvin I. Goldman and Dennis Whitcomb, 71-91. New York: Oxford University Press.

—. 2021. *The Epistemology of Groups.* Oxford: Oxford University Press. doi:10.1093/oso/9780199656608.001.0001.

Landgrebe, Jobst, and Barry Smith. 2023. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear.* New York: Routledge. doi:https://doi.org/10.4324/9781003310105.

Leonard, Nick. 2021. *Epistemological Problems of Testimony.* Edited by Edward N. Zalta. 1 April. https://plato.stanford.edu/archives/sum2021/entries/testimony-episprob/.

Leonard, Nick. 2018. “The Transmission View of Testimony and the Problem of Conflicting Justification.” *American Philosophical Quarterly* 55 (1): 27-35. https://www.jstor.org/stable/45128596.

Lewin, Kurt. 1943. “Forces behind food habits and methods of change.” *Bulletin of the National Research Council*, October: 35-65. doi:https://doi.org/10.17226/9566.

Lewis-Martin, Jimmy. 2022. “What kinds of groups are group agents?” *Synthese* 1-19. doi:https://doi.org/10.1007/s11229-022-03766-z.

List, Christian, and Phillip Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199591565.001.0001.

Littlejohn, Clayton. 2018. “The Right in the Good: A Defense of Teleological Non-Consequentialism.” In *Epistemic Consequentialism*, edited by H. Kristoffer Ahlstrom-Vij and Jeffrey Dunn, 23-47. Oxford University Press. doi:https://doi.org/10.1093/oso/9780198779681.003.0002.

Longo, L., M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, et al. 2024. “Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions.” *Information Fusion* 106. doi:10.1016/j.inffus.2024.102301.

MacCormick, Neil. 1972. “Voluntary Obligations and Normative Powers I.” *Proceedings of the Aristotelian Society, Supplementary Volumes.* Oxford: Oxford University Press. 59-102. doi:https://doi.org/10.1093/aristoteliansupp/46.1.59.

Mathiesen, Kay. 2006. “The Epistemic Features of Group Belief.” *Episteme* 2 (3): 161-175. doi:https://doi.org/10.3366/epi.2005.2.3.161.

McGrath, Sarah. 2011. “Skepticism about Moral Expertise as a Puzzle for Moral Realism.” *Skepticism about Moral Expertise as a Puzzle for Moral Realism* Skepticism about Moral Expertise as a Puzzle for Moral Realism: Skepticism about Moral Expertise as a Puzzle for Moral Realism.

McLeod, Carolyn. 2020. “Damage to Trust.” In *Conscience in Reproductive Health Care: Prioritizing Patient Interests,* by Carolyn McLeod, 65-87. Oxford: Oxford University Press. doi:https://doi.org/10.1093/oso/9780198732723.003.0004.

—. 2002. *Self-trust and reproductive autonomy.* Cambridge: MIT Press. doi:https://doi.org/10.7551/mitpress/6157.001.0001.

McMyler, Benjamin. 2007. “Knowing at Second Hand.” *Inquiry* 50: 511–40. doi:https://doi.org/10.1080/00201740701612390.

—. 2011. *Testimony, Trust ,and Authority.* Oxford: Oxford University Press.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. “A Turing test of whether AI chatbots are behaviorally similar to humans.” *Proceedings of the National Academy of Sciences* 121 (9): 1-8. doi:https://doi.org/10.1073/pnas.2313925121.

Moran, Richard. 2005. “Getting Told and Being Believed.” *Philosophers' Imprint* (Michigan Publishing) 5 (5): 1-29. doi:https://doi.org/10.1093/acprof:oso/9780199276011.003.0013.

—. 2018. *The Exchange of Words: Speech, Testimony, and Intersubjectivity.* Oxford University Press. doi:https://doi.org/10.1093/oso/9780190873325.001.0001.

Morreau, Michael, and Erik J. Olsson. 2022. “Learning from Ranters:The Effect of Information Resistance on the Epistemic Quality of Social Network Deliberation.” In *Social Virtue Epistemology,* edited by Mark Alfano, Colin Klein and Jeroen de Ridder, 553-582. New York: Routledge. doi:10.4324/9780367808952-74.

Mourelatos, Alexander P. D. 1978. “Events, Processes, and States.” *Linguistics and Philosophy* 2 (3): 415-434. doi:https://doi.org/10.1007/bf00149015.

Neil, Mehta. 2016. “Knowledge and Other Norms for Assertion, Action, and Belief: A Teleological Account.” *Philosophy and Phenomenological Research* 93 (3): 681– 705. doi:doi.org/10.1111/phpr.12222.

Nguyen, C. Thi. 2020. “Echo Chambers and Epistemic Bubbles.” *Episteme* 17 (2): 141-161. doi:10.1017/epi.2018.32.

Nguyen, C. Thi. 2022. “Trust as an unquestioning attitude.” *Oxford Studies in Epistemology* 7: 214-244. doi:https://doi.org/10.1093/oso/9780192868978.003.0007.

Nickel, Philip. 2001. “Moral Testimony and its Authority.” *Ethical Theory and Moral Practice* 4: 253–66.

Nozick, Robert. 1981. *Philosophical Explanations.* Harvard University Press).

OpenAI. 2024. *ChatGPT.* [Large language model]. https://chat.openai.com/chat.

—. 2024. *Expanding on how Voice Engine works and our safety research.* 7 june. https://openai.com/index/expanding-on-how-voice-engine-works-and-our-safety-research/.

—. 2023. “GPT-4 Technical Report.” https://cdn.openai.com/papers/gpt-4.pdf.

—. 2024. *Hello GPT-4o.* 13 May. https://openai.com/index/hello-gpt-4o/.

—. 2024. *Video generation models as world simulators.* 15 February. https://openai.com/index/video-generation-models-as-world-simulators/.

Owens, David. 2006. “Testimony and Assertion.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* (Springer) 130 (1): 105-129. http://www.jstor.org/stable/4321791.

Oxford English Dictionary. n.d. *Gatekeeping.* Accessed January 2, 2024. https://www.oed.com/dictionary/gatekeep_v?tab=meaning_and_use#1367883960.

Ólafsson, Ísak Andri. 2023. “Unwanted Knowledge Transmission.” *Synthese* 201 (162). doi:https://doi.org/10.1007/s11229-023-04140-3.

Palermos, Spyridon Orestis. 2022. “Collaborative knowledge: Where the distributed and commitment models merge.” *Synthese* 200: 1-16. doi:https://doi.org/10.1007/s11229-022-03459-7.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet is Hiding from You.* Penguin UK.

Pelling, Charlie. 2014. “Assertion, Telling, and Epistemic Norms.” *Australasian Journal of Philosophy* 92 (2): 335-348. doi:https://www.tandfonline.com/doi/abs/10.1080/00048402.2013.798340.

Plato. 2008. *Meno, by Plato.* Translated by Benjamin Jowett. 21 September. Accessed November 26, 2019. http://www.gutenberg.org/files/1643/1643-h/1643-h.htm.

Prichard, H.A. 2002. “The Obligation to Keep a Promise.” In *Moral Writings*, edited by J. MacAdam, 257–265. Oxford: Oxford University Press. doi:https://doi.org/10.1093/0199250197.003.0012.

Pritchard, Duncan. 2012. “Anti-luck virtue epistemology.” *Journal of Philosophy* 109 (3): 247-279. https://www.jstor.org/stable/43820700.

Pritchard, Duncan. 2015. “Epistemic Dependence.” *Philosophical Perspectives* 29: 305-324. https://www.jstor.org/stable/10.2307/26614571.

Pritchard, Duncan. 2007. "Recent Work on Epistemic Value." *American Philosophical Quarterly* (University of Illinois Press) 44 (2): 85-110. http://www.jstor.org/stable/20464361.

Pritchard, Duncan. 2004. "The Epistemology of Testimony." *Philosophical Issues* (14): 326-348.

Pritchard, Duncan, and John Turri. 2014. *The Value of Knowledge.* Accessed February 17, 2017. https://plato.stanford.edu/entries/knowledge-value/.

Rabinowicz, Wlodek, and Toni Rønnow-Rasmussen. 2000. “A Distinction in Value: Intrinsic and for Its Own Sake.” *Proceedings of the Aristotelian Society.* 33-51. https://www.jstor.org/stable/4545316.

Raz, Joseph. 1990. “Introduction.” In *Authority*, by J. (ed.) Raz, 1-19. New York: New York University Press.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?": Explaining the Predictions of Any Classifier.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco: Association for Computing Machinery. 1135–1144. doi:https://doi.org/10.18653/v1/n16-3020.

Rolin, Kristina. 2015. “Values in science: The case of scientific collaboration.” *Philosophy of Science* 82 (2): 157–177. doi:https://doi.org/10.1086/680522.

Ross, Angus. 1986. “Why Do We Believe What We Are Told?” *Ratio* 38: 69–88.

Roth, Abraham Sesshu. 2017. “Shared Agency.” *Stanford Encyclopedia of Philosophy.* Edited by Edward N. Zalta. 1 May. https://plato.stanford.edu/archives/sum2017/entries/shared-agency/.

Ryan, Mark. 2020. “In AI We Trust: Ethics, Artificial Intelligence, and Reliability.” *Science and Engineering Ethics* 26: 2749–2767 . doi:https://doi.org/10.1007/s11948-020-00228-y.

Rønnow-Rasmussen, Toni, and Michael J. Zimmerman, . 2005. *Recent Work on Intrinsic Value.* Dordrecht: Springer. doi:https://doi.org/10.1007/1-4020-3846-1.

Sato, Katz. 2018. “What makes TPUs fine-tuned for deep learning?” *Google Cloud Blog.* 30 August. https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep-learning.

Scheman, Naomi. 2020. “Trust and Trustworthiness.” In *The Routledge Handbook of Trust and Philosophy,* edited by Judith Simon, 28–40. New York: Routledge. doi:https://doi.org/10.4324/9781315542294-2.

Schmitt, Frederick F. 2010. “The Assurance View of Testimony.” In *Social Epistemology,* edited by Adrian Haddock, Alan Millar and Duncan Pritchard, 216-242. Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199577477.001.0001.

Searle, John R. 1990. “Collective Intentions and Actions.” In *Intentions in Communication.,* edited by Philip R. Cohen, Jerry Morgan and Martha Pollack, 401-415. MIT Press. doi:https://doi.org/10.7551/mitpress/3839.003.0021.

Searle, John R. 1980. “Minds, Brains and Programs.” *Behavioral and Brain Sciences* 3 (3): 417–457. doi:https://doi.org/10.1017/s0140525x00005756.

Searle, John R. 1999. “The Chinese room.” In *The MIT encyclopedia of the cognitive sciences,* edited by R.A. Wilson and F. Keil, 115-116. Cambridge: MIT Press.

Sheikh, H., C. Prins, and E. Schrijvers. 2023. “Artificial Intelligence: Definition and Background.” In *Mission AI,* by H. Sheikh, C. Prins and E. Schrijvers, 15-41. Springer. doi:10.1007/978-3-031-21448-6_2.

Simion, Mona, and Chris Willard-Kyle. Forthcoming. “Trusting AI: explainability vs. trustworthiness.” In *Communication with AI: Philosophical Perspectives,* edited by H. Cappelen and R. Sterken. Oxford University Press.

Simion, Mona, and Christoph Kelp. 2023. “Trustworthy artificial intelligence.” *Asian Journal of Philosophy* 2 (8): 1-12. doi:https://doi.org/10.1007/s44204-023-00063-5.

Simons, Massimiliano. 2022. “Gatekeepers and Gated Communities: The Role of Technology in Our Shifting Reciprocities.” *Philosophy Today 67(2),* 66 (4): 763-779. doi:https://doi.org/10.5840/philtoday202271461.

Simpson, Thomas W. 2012. “What Is Trust?” *Pacific Philosophical Quarterly* 93 (4): 550–569. doi:https://doi.org/10.1111/j.1468-0114.2012.01438.x.

Snyder, D.K., D. Baucom, K.C. Gordon, J.M. Gottman, and W.K. Halford. 2007. *Getting Past the Affair.* Guilford Press.

Sosa, David. 2023. “Belief beyond groups.” *Asian Journal of Philosophy* 1-9. doi:https://doi.org/10.1007/s44204-023-00077-z.

Sosa, Ernest. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, vol. I.* Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199297023.001.0001.

—. 2021. *Epistemic Explanations: A Theory of Telic Normativity, and What it Explains.* Oxford: Oxford University Press. doi:https://doi.org/10.1093/oso/9780198856467.001.0001.

Sosa, Ernest. 2010. “How Competence Matters in Epistemology.” *Philosophical Perspectives* 24: 465-475. doi:https://doi.org/10.1111/j.1520-8583.2010.00200.x.

Swank, Jacqueline M., and Sondra Smith-Adcock. 2014. “Gatekeeping During Admissions: A Survey of Counselor Education Programs.” *Counselor Education and Supervision* 53 (1): 47-61. doi:https://doi.org/10.1002/j.1556-6978.2014.00048.x.

Tollefsen, Deborah Perron. 2015. *Groups as Agents.* Polity.

Tollefsen, Deborah Perron. 2002. “Organizations as True Believers.” *Journal of Social Philosophy* 33 (3): 395-410. doi:https://doi.org/10.1111/0047-2786.00149.

Tuomela, R. 2013. *Social Ontology: Collective Intentionality and Group Agents.* New York: Oxford University. doi:https://doi.org/10.1093/acprof:oso/9780199978267.001.0001.

Turri, John. 2011. “Manifest Failure: The Gettier Problem Solved.” *Philosopher's Imprint* 11: 1-11.

Ulatowski, J. 2022. “Epistemic Gatekeeping, Pride or Prejudice?”

Vendler, Zeno. 1957. “Verbs and Times.” *The Philosophical Review* 66 (2): 143-160. doi:https://doi.org/10.2307/2182371.

Wanderer, Jeremy, and Leo Townsend. 2013. “Is it Rational to Trust?” *Philosophy Compass* 8 (1): 1-14. doi:https://doi.org/10.1111/j.1747-9991.2012.00533.x.

Welbers, Kasper, and Michaël Opgenhaffen. 2018. “Social media gatekeeping: An analysis of the gatekeeping influence of newspapers’ public Facebook pages.” *New Media & Society* 20 (12): 4728-4747. doi:https://doi.org/10.1177/1461444818784302.

Welbourne, Michael. 1983. “A Cognitive Thoroughfare.” *Mind* 92: 410–12. doi:https://doi.org/10.1093/mind/XCII.367.410.

White, David Manning. 1950. “The "gate keeper": A case study in the selection of news.” *Journalism Quarterly* 27 (4): 383–391. doi:10.1177/107769905002700403.

Wright, Stephen. 2016. "The transmission of knowledge and justification." *Synthese* 293-311. doi:https://doi.org/10.1007/s11229-015-0760-y.

Zagzebski, Linda. 2012. *Epistemic authority: a theory of trust, authority and autonomy.* Oxford: Oxford University Press. doi:https://doi.org/10.1093/acprof:oso/9780199936472.001.0001.

Zagzebski, Linda. 2007. “Ethical and Epistemic Egoism and the Ideal of Autonomy.” *Episteme* 4: 252–263. doi:https://doi.org/10.3366/e174236000700007x.

—. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge.* Cambridge: Cambridge University Press.

Zanotti, Giacomo, Mattia Petrolo, Daniele Chiffi, and Viola Schiaffonati. 2024. “Keep trusting! A plea for the notion of Trustworthy AI.” *AI & SOCIETY* 39: 2691–2702. doi:https://doi.org/10.1007/s00146-023-01789-9.

Zhang, Muru, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. “How Language Model Hallucinations Can Snowball.” *ArXiv preprint* 2305, 13534v1. doi:https://doi.org/10.48550/arXiv.2305.13534.

Zhu, Shiqiang, Ting Yu, Tao Xu, Hongyang Chen, Schahram Dustdar, Sylvain Gigan, Deniz Gunduz, et al. 2023. “Intelligent Computing: The Latest Advances,Challenges, and Future.” *Intelligent Computing* 2 (0006): 1-45. Accessed September 9, 2024. doi:10.34133/icomputing.0006.

Zollman, Kevin J.S. 2013. “Network Epistemology: Communication in Epistemic Communities.” *Philosophy Compass* 8 (1): 15-27. doi:https://doi.org/10.1111/j.1747-9991.2012.00534.x .