A Critical Analysis of the Argumentative Theory of Reasoning: Caveats from the Evolution of Human Communication

MSc Thesis (*Afstudeerscriptie*)

written by

Flip Lijnzaad
(born July 26th, 1999 in Cambridge, United Kingdom)

under the supervision of **dr. Karolina Krzyżanowska**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
November 11th, 2024	dr. Aybüke Özgün (chair)
	dr. Karolina Krzyżanowska (supervisor)
	dr. Giorgio Sbardolini
	dr. Marieke Schouwstra

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Contents

Introduction

{ch:introduction}

Communication is one of the most fundamental parts of the human experience: it's hard, nigh impossible, to imagine our lives without it. Compared to our closest evolutionary relatives, our communicative abilities are very sophisticated; human language is incomparable to any non-human animal's system of communication (Cheney and Seyfarth, 1997). One large difference between human and non-human animal communication is humans' ability to communicate more than just the truth-conditional content of their sentences. More often than not, the communicated content of the sentences we utter extends way beyond the words we speak. For example, consider the following exchange (taken from Levinson, 1983, p. 102):

> A: Where's Bill?
> B: There's a yellow VW outside Sue's house.

Taking the utterances at face value, this exchange is quite nonsensical: the utterances seem to have nothing to do with each other. However, this is a completely normal interaction in real life, given our skills for inferring that this yellow VW might belong to Bill, which would point to his current location being Sue's house.

It should be obvious that our abilities for inference play a large role in our communication. This motivates one to ponder on the nature and history of the relationship between our profound communicative abilities and our outstanding inferential skills. Is there some evolutionary reason why we communicate and infer things in the way we do?

Before we continue this thought, a small terminological aside: much of the literature in pragmatics uses the terms 'reasoning' and 'inference' interchangeably (see for example Stephen Levinson's usage of the term 'reasoning' in Levinson, 1983, p. 218). However, in cognitive science a distinction is often made between the two terms (see Mercier and Sperber, 2011, first paragraph). In any case, it should be clear that reasoning and inference are closely related cognitive abilities, and that their (evolutionary) relationship to communication is one worth investigating.

Regarding the relationship between communication and reasoning, cognitive scientists and philosophers Hugo Mercier and Dan Sperber have proposed a revolutionary theory of reasoning (2011), which intended to solve a puzzling

problem regarding the function of human reasoning. It has been a longstanding view in philosophy (due to René Descartes, among others) that the function of reasoning is to improve an individual's knowledge (Schouls, 1972; Walter Jr, 1951). However, in the second half of the twentieth century, influential experimental investigations into humans' reasoning abilities have led some psychologists to conclude that people are not good at reasoning logically (Wason, 1968) or probabilistically (Tversky and Kahneman, 1983). In the decades since these classical experiments, cognitive scientists have proposed various explanations for these findings (e.g. Cosmides, 1989; Hertwig and Gigerenzer, 1999; Oaksford and Chater, 1994). Among these is Mercier and Sperber's *argumentative theory of reasoning* (ATR), which considers the problem from an evolutionary angle: how it could be the case that reasoning has evolved to serve its function poorly?

Mercier and Sperber argue that reasoning actually serves its function well. According to their argumentative theory of reasoning, the main function of reasoning is argumentative; that is, reasoning evolved in humans to facilitate the production of arguments and the evaluation of those of others. In this way, reasoning serves to improve communication, the stability of which is threatened by dishonest communicators. Speakers can argue for their case and addressees can critically evaluate these arguments, and this allows us to move beyond accepting others' testimony merely on trust.

Mercier and Sperber buttress their ATR by formulating testable hypotheses that are entailed by it, and they gather empirical evidence to corroborate these hypotheses. They argue for example that the phenomenon of confirmation bias, usually construed as a flaw of reasoning, makes perfect sense when considering reasoning's argumentative function: if you want to convince your audience, it is reasonable to only seek out evidence that confirms your opinion.

To summarize, in the words of Mercier and Sperber:

> Reasoning has evolved and persisted mainly because it makes human communication more effective and advantageous.
>
> (Mercier and Sperber, 2011, p. 60)

This view, while thought-provoking and intuitively attractive, is certainly not uncontroversial: Mercier and Sperber's theory has been criticized by philosophers and cognitive scientists alike. Critiques range from objections to their characterization of reasoning (Koreň, 2023) to the theory's evolutionary framework (Dutilh Novaes, 2018) and the cognitive-scientific framework underpinning the theory (Chater and Oaksford, 2018; Sterelny, 2018).

This brings us to our current research endeavor. The purpose of this thesis is to critically analyze Mercier and Sperber's argumentative theory of reasoning by evaluating its (implicit) assumptions and scrutinizing its details against the background of the evolution of human communication.

In order to do this, I will first provide some needed context from evolutionary theory. This context will then inform a comprehensive analysis of human communication from an evolutionary perspective. This analysis will subsequently constitute the foundations for my criticism of the ATR. After dissecting

the ATR in detail, we are then able to assess the plausibility of its component parts and come to a conclusion on the theory's tenability.

The consideration and combination of different disciplines, from epistemology to psychology of reasoning and from the philosophy of biology to evolutionary anthropology, situates this thesis as an endeavor in *synthetic philosophy*, as described by Eric Schliesser:

> [synthetic philosophy is] a style of philosophy that brings together insights, knowledge, and arguments from the special sciences with the aim to offer a coherent account of complex systems and connect these to a wider culture or other philosophical projects (or both). (Schliesser, 2019, pp. 1–2)

This thesis will provide constructive criticism for Mercier and Sperber's argumentative theory of reasoning. It is, however, way beyond the scope of this thesis to propose an alternative theory to the ATR; I will highlight problems with the theory and propose possible ways to fix these problems, but I conjecture that much additional philosophical and/or empirical work is needed to 'fix' the argumentative theory of reasoning.

The structure of this work is as follows. Chapter 1 provides some necessary background on evolution. In this chapter I will discuss evolutionary causation, provide context about evolutionary psychology, and discuss terminological issues associated with function in evolution. This chapter concludes by laying out a methodology for the next chapter. Chapter 2 analyzes human communication from an evolutionary perspective, following the methodological steps outlined in the previous chapter. In particular, this chapter will see us discussing the function of communication at length, which will provide us with important anchor points for scrutinizing the ATR. In Chapter 3, I will expound the argumentative theory of reasoning by discussing Mercier and Sperber (2011) and a number of foundational papers that preceded this seminal paper. All of this then culminates in Chapter 4 with the critical analysis of the ATR, combining the findings from Chapter 2 on the evolution of communication with the details and implicit assumptions of the ATR that became apparent in Chapter 3. Moreover, I critically examine a key concept of the ATR, *epistemic vigilance*, by discussing a critical response to Sperber's work and his reply to it.

Finally, Section 4.4 recounts the whole story and highlights avenues to continue this research.

1 | The chicken and the egg: the evolutionary approach

{ch:evolution}

Before we immerse ourselves in the evolutionary perspective integral to the argumentative theory of reasoning, we should start by laying out some groundwork on evolution. This chapter by no means intends to provide a comprehensive overview of the issues in evolutionary theory, since this is a vast field of research in its own right with widely diverging opinions on a number of specifics of the process of evolution[1]. The purpose of this chapter is instead to provide a starting point for our analysis of the evolution of human communication, against which the ATR can subsequently be criticized.

In this chapter, we will consider the processes underlying evolution, and discuss causation in evolution. Moreover, we will touch on some considerations surrounding the evolutionary approach to human psychology, and discuss the use of teleological terminology in evolutionary theory. Lastly, this chapter will see us outlining a methodology for the investigations of Chapter 2.

1.1 Evolution: biological vs. cultural

{sec:evo-bio-culture}

Although the term 'evolution' is in everyday usage most commonly interpreted as 'Darwinian evolution', or 'natural selection', the concept of evolution can be stripped down to a very broadly construed version, which may prove to be illuminating.

In general, any process of selection can be taken to consist of three steps: (1) variation, (2) sorting[2], and (3) retention (Donahoe, 2003). I will first discuss how each of the steps of this process are construed in standard evolutionary theory, and then how the process of *cultural evolution* fits this conception of selection.

Firstly, in standard evolutionary theory the processes that introduce variation are mutation and migration. Mutation concerns the changes in an or-

[1] See for example Ariew et al. (2002) and Uller and Laland (2019) for an overview of topics in evolutionary theory and evolutionary causation.

[2] Donahoe (2003) uses the term 'selection' for this step. I follow Heyes (2018) and Scott-Phillips et al. (2013) in using the term 'sorting', to avoid confusion with the full process of selection (which consists of the three steps outlined here).

ganism's DNA, and migration concerns the movement of (genetic material of) organisms from one population to another (Scott-Phillips et al., 2013).

Secondly, in standard evolutionary theory the sorting process amounts to natural selection, as well as genetic drift. Here we will only focus on the former of these two processes[3]. The sorting process of natural selection[4] favors variants that enhance an organism's fitness and disfavors variants that diminish its fitness. An organism's fitness corresponds to how likely they are to leave offspring in the next generation compared to organisms with a different genetic makeup. This fitness relates to both the organism's chances of survival as well as its chances of reproducing.

Thirdly, in standard evolutionary theory, the process responsible for retention is genetic inheritance, i.e. the transmission of characteristics from parents to their offspring through the genetic material (DNA) parents impart on their offspring. As a result of these three steps of the selection process then, the genes that enhance an organism's fitness persist over time in the population, resulting in adaptation (Scott-Phillips et al., 2013).

Now, the evolution of *culture* can also be said to operate by these principles (Heyes, 2018). In this context, culture is understood at its core as *information*; more specifically, it is information "that we inherit from others through social interaction (via certain kinds of social learning)" (Heyes, 2018, p. 30). Let us now consider the cultural processes that constitute each of the three steps of the selection process, following Heyes (2018, pp. 33–34).

Firstly, variation in culture is introduced by error and by innovation. For example, an individual might secure their fishing line with four instead of three knots by accident; or they may produce this variation intentionally, in an effort to improve upon the practice. Secondly, the sorting of behaviors and habits in culture can happen through two different routes. A behavior can be favored because of some property inherent to the behavior that makes it "more noticeable, learnable, or memorable than others" (Heyes, 2018, p. 34) and thus more likely to be copied. For example, the four-knot fishing line might be favored if it is easier to construe than the three-knot version. Also, a behavior or habit may be favored in a more 'classic' evolutionary way: a habit may be favored because it improves the fitness of the individual, such that individuals with that habit are more likely to survive and reproduce than individuals with an alternative habit. If the four-knot fishing line makes people more successful in catching fish, this practice increases their chances of survival and is thus favored. Thirdly, culture may be retained through cultural inheritance: cultural practices are passed on between people through mechanisms of social learning.

Whether or not the process of cultural evolution can be taken to be analogous to that of biological evolution, is not an uncontroversial issue; see Claidière

[3]In a nutshell, genetic drift can be described as a probabilistic sorting process, resulting from a sampling error due to populations being finite in size. However, genetic drift is a hotly debated concept in evolutionary theory (see Millstein, 2021, for an overview), so I mention it here only for the sake of completeness.

[4]Not to be confused with the full three-step process, which is (confusingly) usually also referred to as natural selection.

et al. (2014) and Stanley (2021) for discussion. In this thesis however, I will follow Heyes (2018) in the assumption that cultural evolution is 'Darwinian'; that is, it abides by mechanisms that are analogous to those of biological evolution. As a consequence, in discussing how human communication and reasoning evolved, we may remain agnostic on the kind of evolution responsible for this evolution, since the mechanisms underlying biological and cultural evolution are assumed to be analogous to each other. Thus, for the remainder of this thesis, I will use the term 'evolution' to talk about the development of traits over time, remaining agnostic about whether this development is due to biological evolution or cultural evolution.

Let me conclude this discussion of the workings of evolution with a brief detour to the field of evolutionary game theory, which (as the name implies) applies game theory to the evolution of animal behavior. As we will see in Chapter 3, Mercier and Sperber's theory draws heavily on game-theoretic notions and concepts, in particular on *cost-benefit analyses* (see Sperber, 2001; Sperber et al., 2010). In general, one may analyze an animal's behavior in terms of the costs and benefits the behavior yields to the animal. For example, hunting down a large prey animal has its costs (it takes energy and effort) and its benefits (it yields sustenance). The payoff of the behavior can then be considered to act as a proxy for the fitness of the animal: "[payoffs] are meant to represent how much the outcome of the game increases fitness" (McNamara and Weissing, 2010, p. 118)[5]. For example, in the case of hunting game, if the behavior's benefits exceed its costs, it can be taken to increase the hunter's fitness.

Add wrap-up sentence

1.2 Causation in evolution

{sec:causation-evolution}

Next, we will dip our toes into the waters of causation in evolution. As it turns out, causation in evolution is not a simple notion; consider for example a moth whose wings provide it with camouflage due to their coloration. The camouflage of the wings is an *effect* of their coloration; yet, it is precisely the camouflage the coloration provides that is the *cause* of the coloration being present at all (Lipton, 2009).

In evolutionary causation, one may distinguish *proximate* from *ultimate* causes (Mayr, 1961). Proximate causes are the immediate influences on a trait: they explain how the trait results from the internal and external factors causing it. For example, the proximate cause of the coloration of the moth's wings would be the biochemical processes during the moth's development that result in that particular pattern of colors. Ultimate causes, on the other hand, provide the higher-level historical and evolutionary explanation of those traits. In the case of the moth, the ultimate cause for the coloration of its wings is the enhanced fitness of the moth due to the camouflaging properties of the coloration (Lipton, 2009). In other words, these two different causes relate to two different explana-

[5] McNamara and Weissing (2010, §4.4) provide some caveats to this conception of payoff as proxy for fitness. However, I consider a comprehensive discussion of the game-theoretical approach to evolution to be out of scope of this thesis.

tory questions: the proximate cause is related to the *how*-question (*how* did a trait come about?), whereas the ultimate cause is related to the *why*-question (*why* did a trait come about?). According to Mayr (1961), who pioneered the distinction[6], one needs to answer both of these explanatory questions in order to obtain a complete understanding of a trait.

In a seminal proposal considered by some to be an extension of Mayr's dichotomy (Laland et al., 2013), Nico Tinbergen (1963) outlined four questions central to the study of animal behavior. In order to fully understand a pattern of behavior, he argued, one must consider (1) the behavior's proximate causation, (2) its survival value, (3) its lifetime development, and (4) its evolutionary history. Since Tinbergen's proposal, other authors have grouped these four questions according to Mayr's proximate-ultimate distinction, characterizing the 'causation' and 'development' questions as proximate questions (*how*-questions) and the 'survival value' and 'evolutionary history' questions as ultimate questions (*why*-questions) (Bateson and Laland, 2013). Tinbergen's framework has had an extensive and lasting influence on the study of animal behavior, and his questions continue to be used by biologists to this day (Bateson and Laland, 2013). Hence, I will let this framework guide our investigation into the evolution of human communication; we will flesh this out in Section 1.5.

1.3 Evolutionary psychology

{sec:evol-psych}

In order to get a grip on the evolution of cognitive capacities such as reasoning and communication, let us first zoom out to consider the field of evolutionary psychology as a whole. What is the merit, and the validity, of adopting an evolutionary approach to human psychology?

The field of evolutionary psychology concerns itself with trying to understand human behavior using evolutionary theory, by looking into the past and considering how our ancestors must have adapted to their environment in order to survive and reproduce. Researchers in the social sciences and humanities have historically been wary of using evolutionary approaches to study human behavior, because evolutionary theory has been abused for prejudiced ends in the past[7]. Moreover, evolutionary-psychological research has received the criticism that too much of it is "just-so" storytelling and post-hoc explanation of known phenomena, sometimes accompanied by a sensationalist spin on the story (Laland and Brown, 2002). However, if these pitfalls are avoided, looking at human psychology from an evolutionary perspective can be an illuminating endeavor. Let us now consider some of the central concepts and assumptions of evolutionary psychology.

In order to explain humans' psychological mechanisms, evolutionary psychologists look to the concept of an *environment of evolutionary adaptedness* (EEA). The EEA is the environment in which these psychological mechanisms must have come into being; usually the EEA is identified as hunting and gathering

[6] See Laland et al. (2013, p. 720) for a discussion of the exact origins of the distinction.
[7] See Laland and Brown (2002, pp. 19–20) for an overview.

groups on the African savannah in the second half of the Pleistocene, between 1.7 million and ten thousand years ago (Laland and Brown, 2002). There are two key assumptions underlying the use of the concept of an EEA. The first is that our modern-day environment is too different from that of our ancestors for us to use it to explain why and how our psychological mechanisms evolved in the past. The second assumption is that for our psychological mechanisms to be as complex as they are, they must have evolved slowly; because of this, they must have evolved a considerable amount of time ago without changing significantly since the Stone Age.

There are a number of issues associated with the use of the concept of the EEA (Laland and Brown, 2002). Firstly, we do not know very much about the environment of our ancestors, so the specifics of the EEA may be filled in as is seen fit for one's purposes. Secondly, it may be argued that we do not know enough about the process of evolution to make the second assumption. While evolution does in general operate on a large timescale, there is empirical evidence that the process can also be faster, operating on a timescale of thousands of years, or less than 100 generations (Laland and Brown, 2002, pp. 190–191 and references therein). Lastly, the argument can be made that for our species to have flourished and dominated in the way that it did, we must have remained adaptive to our changing modern environments after the Stone Age.

Despite the issues associated with the concept of the EEA, it is instrumentally valuable in reminding us to consider the state of the environment and its role in the evolutionary process. For the purposes of this thesis, we need not commit to any strong assumptions about the nature and properties of the EEA. The most important assumption I will make, following influential scholars like Michael Tomasello (2009), is that humans throughout history have been dependent on cooperation and strong social groups for their survival. In the EEA, humans lived together in groups and relied on hunting game and gathering plants for their nutrition. In this lifestyle, cooperation is a "necessary element of human life" (Apicella and Silk, 2019, p. R448) in a number of ways. Firstly, hunting is a 'high risk, high reward' endeavor: the returns are variable, but when hunting does succeed the yield is often large – sometimes even too large for the hunter and his relatives. In this case, food sharing within or between groups is beneficial. Another way in which early humans counterbalanced the variable returns of hunting was by also relying on gathering plant foods, which yielded more predictable returns. In this case cooperation through shared labor was also beneficial, since some foods required complex foraging techniques or complex processing (e.g. cooking) before consumption. Lastly, cooperation in early humans manifested itself in 'cooperative breeding', where the responsibilities of childcare are spread among multiple caregivers. Moreover, mothers and children relied on the efforts of others for their food (Apicella and Silk, 2019). We will discuss human cooperation in more detail in Section 2.4.1, in particular how its evolution relates to the evolution of human communication.

1.4 Teleological notions in evolutionary theory

{sec:teleology}

Next, let us touch on a terminological issue that deserves some discussion. Biological literature frequently makes use of *teleological* terminology, that is, terminology that implies goal-directedness of the processes it describes. Such terminology includes concepts like the *design* of a trait, and the *function* or *purpose* of a trait. This teleological terminology has its roots in pre-Darwinian conceptions of nature: it originates from Aristotle's views on nature, and was subsequently adopted by creationist Muslim and Christian scholars (Johnson, 2005). At first glance, the usage of these terms in discussing evolution would thus seem to be inappropriate. Evolution is a process of nature, not purposefully performed by an agent, and it is thus without any intentionality or goals. Indeed, teleological explanations in biology are somewhat controversial (see Ayala, 1999, p. 27 for discussion). However, explanations in terms of goals and functions have considerable *instrumental* value in describing evolutionary processes. In light of this, consider the following characterization of teleological explanations:

> Teleological explanations account for the existence of a certain feature in a system by demonstrating the feature's contribution to a specific property or state of the system, in such a way that this contribution is *the reason why the feature or behaviour exists at all*. (Ayala, 1999, p. 13)

In this respect, the evolutionary process of adaptation merits a teleological explanation. A trait's survival value – its 'contribution to a specific property or state of the system' – is the reason that the trait has persisted throughout evolution. Returning to the moth's colored wings, their function is to provide camouflage, and this contribution to the moth's survival is the reason the coloration exists in the first place.

The distinction between proximate and ultimate causes we saw in Section 1.2 can be applied to teleological explanation as well, yielding the distinction between proximate and ultimate *ends* of features (Ayala, 1999, p. 18). The proximate end is then the 'immediate' function the feature serves (e.g. the camouflage), and the ultimate end is the reproductive success of the organism.

We can also see this in Mercier and Sperber's view that the function of reasoning is to make communication "more effective and advantageous" (Mercier and Sperber, 2011, p. 60). According to this view, the improvement of communication is the proximate end of reasoning. Inherent to this view is then the assumption that improving communication contributes to humans' fitness. Therefore, in order to critically analyze Mercier and Sperber's argumentative theory of reasoning, we should first investigate human communication from an evolutionary perspective.

1.5 Adopting and adapting Tinbergen's four questions

{sec:tinbergen}

As mentioned in Section 1.2, Tinbergen (1963) proposed an influential framework of problems[8] that should be addressed if one intends to give a complete account of a behavior an animal exhibits. The four problems that Tinbergen argued to be central to the study of behavior are a behavior's proximate causation, survival value, lifetime development, and evolutionary history. Although these four problems were originally introduced with regards to animal behavior, the framework has since been widely adopted for analyzing any trait of an organism (Bateson and Laland, 2013)[9]. Let me now discuss each of these problems in more detail, such that we can ultimately come to a set of questions to guide us in investigating the evolution of human communication in Chapter 2.

The first Tinbergen problem is that of the 'mechanistic' causation of the behavior; in other words, the proximate causation of the behavior. In our case, addressing this problem would entail a detailed investigation of the neurological processes underlying communicative behaviors. This problem, however interesting, will not be addressed in this thesis. The reason for this is that more empirical and conceptual research would be necessary in order to give a satisfactory account of the exact neurological processes underlying human communication. Although Tinbergen (1963) emphasized that we can only gain a full understanding of a behavior if all four problems are addressed simultaneously (see also Bateson and Laland, 2013), I believe I am justified in considering the proximate-causation problem to be out of scope for the current endeavor.

The second problem that Tinbergen outlines, relates to how the behavior contributes to the chances of the animal surviving. This survival value is, in teleological terms, the function of the behavior. However, the use of the term 'function' may obscure the fact that a trait's function can change over time: the *current* utility that a trait has, may not be the same as the *original* utility it had (Bateson and Laland, 2013). For example, feathers originally evolved for temperature regulation in the evolutionary predecessors of birds, and were later adapted for flight (Bateson and Laland, 2013; Benton et al., 2019). Another example of the current and original utilities of a trait not lining up, is fat retention in humans. Our ability to store energy in the form of fat originally contributed to our survival value by providing a buffer against malnutrition and fluctuating energy supplies (Wells, 2006). In light of the obesity epidemics in present-day western societies however, this trait can hardly be considered to positively contribute to our chances of survival. In our investigation in Chapter 2, we will focus on the *original* utility of communication. Since we are interested in how human communication evolved and how the evolution of reasoning relates to

[8] In the literature (e.g. Bateson and Laland, 2013), the terms 'problems' and 'questions' are used interchangeably. I will take 'problems to address' and 'questions to answer' to be synonymous, and will use these terms interchangeably.

[9] It can even be used to gain understanding of nonliving systems, such as traffic lights (Bateson and Laland, 2013).

this, only the original utility of communication is relevant here.

As we saw in Section 1.1, an organism's fitness is not only determined by their chances of *survival*, but also their chances of *reproducing*. As a consequence of this, the survival value a trait brings to an organism is not the only reason why the trait may persist throughout evolution. A trait is also more likely to appear in future generations if it improves an organism's chances of reproducing. In the methodological framework proposed here I will amend Tinbergen's question on survival value by broadly speaking of the *utility* of a trait, which then denotes the way the trait contributes in general to the fitness of the organism.

The third question that is essential for gaining understanding about a behavior is the question of how the behavior emerges and changes throughout the development (ontogeny) of the animal.

Add one sentence

The fourth and last problem considered by Tinbergen is that of the evolutionary history of the behavior: in order to provide a complete explanation of a behavior, one must look at how it evolved throughout history. To form hypotheses about this, one must look to whether and how the behavior presents itself in the closest evolutionary relatives of the animal. Bateson and Laland (2013) maintain that for traits related to human cognition, this question about evolutionary history should be split up into two questions. They argue that, due to the influence of not only biological evolution but also culture on the development of a trait, one should distinguish two kinds of evolutionary history, leading to the questions "Which historical processes were responsible for the [trait]?" and "How can its trajectory be explained?" (Bateson and Laland, 2013, p. 714). However, as I concluded in Section 1.1, we are justified in remaining agnostic about which processes were responsible. Therefore, we will only take up the latter of these two questions and ask ourselves how the evolutionary trajectory of human communication can be explained.

Lastly, a problem that is mentioned by Tinbergen in his original paper (1963) but is not included as one of the core problems of his framework[10], is the problem of *describing* the observed behavior. In the case of describing communication, this issue is akin to the problem of defining and delineating what we take to be communication. This is by no means a trivial issue, which is why I will include it in the set of questions we will ask ourselves in Chapter 2.

1.6 Conclusion

{sec:evo-conclusion}

Throughout this chapter, it has become apparent that for many of the topics in evolutionary theory discussed here, there exists no consensus among its practitioners. Since the purpose of this thesis will not be to provide a complete causal framework for the evolution of reasoning and communication, we will be able to cast aside some of the issues plaguing the frameworks discussed in this chapter. We will proceed cautiously, using the concepts outlined without needing

[10] And more or less never included by authors discussing his framework (Allen and Bekoff, 1995; Laland and Brown, 2002; Laland et al., 2013).

to account in detail for their shortcomings. Let us conclude this chapter by first gathering three key assumptions that will inform this thesis, and then formulating the methodological questions that will guide our investigations.

The first is the assumption that we are justified in wanting to explore human reasoning and communication from the perspective of evolution. Despite some of the issues raised against evolutionary psychology as a field of study (Laland and Brown, 2002), it cannot be denied that reasoning and communication are cognitive capacities that must have emerged somewhere on our evolutionary journey, through processes of selection (i.e. as a result of variation, sorting and retention).

The second assumption is that in our analysis we may remain agnostic about whether biological or cultural evolution is responsible for the emergence of reasoning and communication, since these types of evolution have analogous underlying mechanisms of selection.

The third assumption is that throughout our evolutionary trajectory, we have been dependent on fellow humans for our survival, relying on cooperation and strong social groups.

Finally, in Section 1.5 we adapted Tinbergen's (1963) framework to arrive at four questions that will guide our analysis in Chapter 2. These questions may be formulated as follows (presented in the order in which we will discuss them):

Definition	How can communication be defined and delineated?
Evolution	What is the evolutionary history of human communication: how can its evolutionary trajectory from our nearest evolutionary relatives to us be explained?
Development	How does communication develop throughout childhood?
Utility	What is the original utility of communication to humans?

2 | The evolution of human communication

{ch:communication}

Human communication is uniquely sophisticated and remarkably pervasive. In this chapter, we will dive head-first into studying its evolution, guided by the four methodological questions expounded in Section 1.6. First, we will consider different conceptions and definitions of communication in Section 2.1. Then, we will consider how non-human animals and human children communicate, respectively, in Sections 2.2 and 2.3. Lastly, we will discuss the utility of communication at length in Section 2.4. In doing so, we will also consider the evolution of human cooperation, the issue of the stability of communication, and the role deception plays in this.

2.1 Conceptions of communication

{sec:comm:definition}

There are many different ways in which organisms may communicate with each other, and indeed many different ways in which one may define communication. In any case, communication is a process necessarily involving a sender (communicator) and at least one receiver (addressee). In the literature, the terms 'communication' and 'signaling' are used interchangeably; therefore, I will consider them to be interchangeable as well.

An influential conception of communication, due to biologists Dawkins and Krebs (1978), maintains that communication occurs "when an animal, the actor, does something which appears to be the result of selection to influence the sense organs of another animal, the reactor, so that the reactor's behavior changes to the advantage of the actor" (Dawkins and Krebs, 1978, p. 283). In other words, communication happens when an animal does something to influence another animal's senses so as to change their behavior to the sender's benefit. Embedded in this conception is an evolutionary component: the communicative behavior "appears to be the result of selection".

A similar, but slightly more general conception of communication is explicated by Freeberg et al. (2019) in their discussion of communication as it relates to social cognition. They define communication as follows:

> Communication involves an action or characteristic of one individual that

> influences the behaviour, behavioural tendency or physiology of at least one other individual in a fashion typically adaptive to both (p. 281)

Compared to Dawkins and Krebs's definition, this definition widens the scope of communicators from animals to organisms in general: "All living things communicate" (Freeberg et al., 2019, p. 281). Moreover, this definition includes the possibility of communicating to multiple receivers. The last notable difference between the two definitions is that on the former definition, communication necessarily benefits the sender, whereas on the latter, communication typically benefits both the sender and the receiver. This mutual benefit is an important aspect of communication, as we will see in Section 2.4.2.

This definition of communication due to Freeberg et al. (2019) not only covers our case at hand – humans speaking and listening to each other – but also encompasses cases such as for example bacteria communicating population density through chemical signals (Federle and Bassler, 2003), squirrels producing alarm calls to alert their relatives of danger (Sherman, 1977), and plants communicating distress through airborne sounds (Khait et al., 2023). A particular example of communication that we will return to in Section 2.4.2 is that of the peacock's tail. Intuitively, it is perhaps not immediately obvious that this constitutes communication. However, one may think of it as the tail 'sending a message' to peahens: it signals the peacock's fitness to them. Following the above definition, the peacock's tail constitutes communication because it is a characteristic that influences the behavior of peahens in a fashion adaptive to the peacock (because it improves his chances of reproducing) as well as the peahens (because it allows them to choose a quality male for mating).

Communication may be modeled in two fairly different ways, using the classical *code model* of communication, and the *ostensive-inferential* model of communication (Scott-Phillips, 2015). In the former model, communication involves processes of coding and decoding messages. The coding, on the side of the sender, involves a mapping between the state of the world and a behavior (the behavior being the signal they send). Then decoding, on the side of the receiver, involves a mapping between two behaviors: the signal sent by the sender, and the subsequent response of the receiver. If the mappings are properly calibrated to each other, communication between sender and receiver can be said to have occurred (Scott-Phillips, 2015).

The code model is considered to be inadequate for accurately describing human communication, because it fails to account for the *underdeterminacy of meaning*: merely taking the message at surface value, one cannot account for the meaning[1] that the message conveys to the sender (Scott-Phillips (2018); cf. Quine (1960)'s views on radical translation). The *ostensive-inferential model* of communication, introduced by Sperber and Wilson (1986) as part of their neo-Gricean framework of relevance theory, deals with this by taking into account the intentionality inherent in human communication.

Add example

[1] I use the intuitive, colloquial interpretation of the term 'meaning', as I consider further explorations into the philosophy of language out of scope for this thesis.

In the ostensive-inferential model, one may speak of a sender's *informative intention*, which is their intending for the receiver to believe something. The sender's *communicative intention* is then their intending for the receiver to believe that they have an informative intention. The sender may then express or convey this communicative intention to the receiver with an *ostensive* behavior. If the receiver receives their communicative intention, then ostensive-inferential communication has occurred.

Currently, there is no evidence that any species other than humans communicate ostensively (Scott-Phillips, 2018). Consequently, one may conceptualize the code model and the ostensive-inferential model as capturing two different types of communication. The code model then captures the way that non-human animals communicate, and the ostensive-inferential model then captures the way that humans communicate with each other.

Finally, it is worth noting that Mercier nor Sperber ever defines what they take as the definition of communication. I will assume that they adhere to the ostensive-inferential conception of communication, since this theory was co-introduced by Sperber himself (Sperber and Wilson, 1986) and Mercier and Sperber often refer to the principles of relevance theory in their writings (Mercier and Sperber, 2009, 2011; Sperber et al., 2010).

2.2 Communication in non-human animals

{sec:comm:phylogeny}

Now we turn to the methodological question on evolutionary history, and have a look at communication in non-human animals.

Communication is used by non-human animals for a wide range of purposes, and it can be elicited by a number of different stimuli. Moreover, communicative behaviors can manifest themselves in different modalities: not only can animals communicate through vocalizations, they may also communicate through gestures or glances. For example, a chimpanzee can communicate to another chimpanzee that it wants to play with them by slapping the ground and directing their gaze at them (Call and Tomasello, 2007).

One can broadly distinguish between communication in aggressive and cooperative interactions (Seyfarth and Cheney, 2003). In aggressive interactions, primates may for example use communication in order to intimidate, by signaling their size and willingness to fight. This minimizes the chances of a physical altercation or fight actually happening, which minimizes the chance of injury for both the dominant and the subordinate animal. In cooperative interactions on the other hand, where the interests of the signaler and the receiver overlap, communication can be used to alert others about predators, to coordinate foraging activities and to facilitate social interactions. For example, animals may communicate to each other "information about predators or the urgency of a predator's approach, group movements, intergroup interactions, or the identities of individuals involved in social events" (Seyfarth and Cheney, 2003, p. 168).

In animals in general, vocalizations are most often elicited not by just one stimulus, but rather a complex combination of them. Moreover, the "history

of interactions between the individuals involved" (Seyfarth and Cheney, 2003, p. 151) can also play a role in eliciting vocalizations. As for the 'immediate' stimuli eliciting vocalizations, we may distinguish between sensory stimuli on the one hand and mental stimuli on the other. Sensory stimuli are to stimuli received through the external senses, such as visual, auditory and olfactory senses. For example, if a male diana monkey sees a leopard (sensory stimulus), he will produce an alarm call, alerting his conspecifics of the presence of the leopard as well as signaling to the leopard that it has been detected (communicative behavior) (Zuberbühler et al., 1997). Mental stimuli on the other hand can be viewed as the mental states one attributes to another individual. The latter type of stimulus elicits the majority of vocalizations in human conversation, but there is no evidence that the attribution of mental states to others causes vocalizations in other animals (Seyfarth and Cheney, 2003). Moreover, there is a large debate among comparative psychologists and philosophers about whether non-human animals possess the ability to attribute mental states to others (i.e., have a theory of mind) at all; see Andrews (2015, Chapter 6) for discussion.

As we will discuss deception in communication a great deal (e.g. in Section 2.4.3), let us briefly touch on deception in non-human animals as well. Primates possess the ability to deceive others ; however, compared to humans, their ability to deceive is rather primitive (Dor, 2017). One argument for this is that more sophisticated deception requires abilities for theory of mind – as (Lee, 2013) puts it, "Lying in essence is ToM in action" (p. 91)" – and primates do not possess these abilities to the extent that humans do (cf. Andrews, 2015, Chapter 6). The difference between human and primate deception also lies in the fact that language uniquely enables humans to communicate with each others' imaginations (Dor, 2017). This opens up countless avenues for deception by the fabrication of stories. For primates on the other hand, without language, it is difficult to fabricate stories to deceive others. They can hide information: for example, they might suppress the food call they would usually expel upon finding food in order to keep this food for themselves. However, their ability to fabricate information is very limited (Dor, 2017).

Elaborate

2.3 Communication in children's development

{sec:comm:ontogeny}

Let us now turn our attention from non-human animals to human children and consider how children communicate throughout their development. Children first start communicating ostensively through pointing gestures at around 9–10 months of age (Carpenter et al., 1998, pp. 20–21 and references therein). Although at first glance, pointing may seem like a simple behavior, it may be used in a number of communicative contexts to convey a wide range of messages and intentions. For example, infants may point at a cup to indicate that they want to drink from it (i.e., pointing to request), but they may also point to a hidden object that their parent is searching for (i.e. pointing to inform).

Pointing can serve either of two communicative motives: an *imperative* mo-

tive, where the pointer requests something from someone, and a *declarative* motive, where the pointer shares their experiences and emotions with someone. This basic account of pointing motives can be extended upon by distinguishing between declaratives as *expressives* (sharing attitudes and emotions) and declaratives as *informatives* (providing information). Furthermore, one can conceive of imperatives as a continuum, with the underlying motive ranging on a scale from individualistic – e.g. forcing someone to do something – to cooperative, e.g. indirectly making a request to someone by informing them of some desire (Tomasello, 2008). Primates only use pointing gestures imperatively, not declaratively (Gómez, 2004; Tomasello, 2005). In particular, Bullinger et al. (2011) found that primates only used a pointing gesture when they benefited from this act of communication, while 25-month-old infants pointed regardless of whether they themself benefited from this action. As Tomasello (2009) concludes, "[the infants] could not help but be informative" (p. 17).

Tomasello (2008) hypothesizes that in order to communicate intentionally, like children begin doing around their first birthday, first the skills and motivations for *shared intentionality* need to be present in the infant. Shared intentionality is the "ability to participate with others in interactions involving joint goals, intentions, and attention" (Tomasello, 2008, p. 139). Communicative pointing behaviors in infants emerge around the same time as skills and motivations of shared intentionality do, which according to Tomasello confirms this hypothesis of dependency between them.

Tomasello further investigates what he calls *pantomiming* or *iconic gestures*, which are symbolic or representational gestures. He presents empirical evidence that these kinds of gestures rely heavily on convention for their meaning, and that the acquisition and usage of these conventions bears a strong resemblance to the acquisition and usage of language. Moreover, in the early stages of language production, toddlers communicate cross-modally by combining gestures with words to form propositions (Iverson and Goldin-Meadow, 2005). In this way, the development of communicative abilities in children can be considered to start in the gestural modality, with language emerging later on (Tomasello, 2008).

Add example

Add evidence

In short, infants first acquire the skills and motivations needed for shared intentionality; then they acquire the skills and motivations for communicative pointing; and then they acquire the ability to use iconic gestures and language around the same time.

Let us, like in Section 2.2, touch briefly upon deception in ontogeny. Lee (2013) notes that there are two components to lying as a speech act. On the one hand, lying is governed by intentionality, and as such it requires the ability to represent mental states of oneself and others (i.e. theory of mind). On the other hand, lying is governed by conventionality, and as such it requires knowledge about social norms against lying. Both of these components are acquired over the course of childhood, in particular in the first four years of development (Lee, 2013); as such, lying can be considered to be a skill that has to be acquired (Meibauer, 2018).

2.4 The utility of communication

{sec:comm:function}

Finally and arguably most importantly for our endeavor, let us have a look at the utility, or function, of communication.

Essentially, communication facilitates interaction between individuals. This interaction may be either cooperative or competitive in nature, as we have seen in Section 2.2 when discussing non-human animal communication. Whether the communicative event is cooperative or competitive in nature depends on the interests of the interlocutors. If the interests of interlocutors overlap or align, their communication can be considered to be cooperative. For example, if two individuals engage in collaborative hunting of a large prey animal, their interests (catching the prey together and sharing it) align and they will thus use communication for cooperative purposes – i.e., to coordinate their hunting activity. On the other hand, if interlocutors' interests do not overlap, or they even oppose each other, communication can be considered to be competitive. For example, if two individuals compete for a smaller prey animal, their interests (catching the prey by themselves and keeping it for themselves) oppose each other, and their communication would thus be competitive. This competitive communication could for example entail verbal intimidation, which may be evolutionarily more advantageous than physical intimidation (i.e. fight) because of a reduced risk of injury.

As argued by Tomasello (2008, 2009), the cooperative setting constitutes the 'birthplace' of the unique features of human communication, and the competitive use of human-style communication must have emerged later:

> The use of skills of cooperative communication outside of collaborative activities (e.g., for lying), came only later.
>
> (Tomasello, 2008, p. 325)

In particular, it has been argued that language could only have emerged in cooperative settings (Dor, 2017; Tomasello, 2008). This is because for the complexity of language to arise, more frequent and prolonged interactions are necessary (Benítez-Burraco et al., 2021). As Dor (2017) writes,

> The collective effort of the invention and stabilization of the new technology [namely, language] must have been based on high levels of reliability and trust between the inventors: otherwise, indeed, they would not have been able to get the system going. (p. 50)

We return to a discussion of reliability, trust and this idea of 'getting the system going' in Section 2.4.2. This discussion of cooperativeness naturally connects to pragmatics; more specifically, it will be illuminating to consider how the aforementioned cooperative function of communication relates to Grice's cooperative principle:

> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. (Grice, 1975, p. 45)

It appears that cooperativeness in a Gricean sense is distinct from the cooperativeness we have been talking about so far. Regarding this, Dor (2017) notes that 'honesty' as a concept might be interpreted in two ways. On the one hand, one can consider the honesty of a signal to be its truthfulness: an honest signal is a true signal, and a deceitful signal is a false signal. Alternatively however, one may conceptualize the honesty of a signal not as its truthfulness, but rather as its 'helpfulness'. In this interpretation, the honesty of a signal corresponds to the benefits and costs the sender and receiver incur as a result of the communicated signal. In the case of animals communicating, these two conceptions of honesty might very well coincide – i.e., truthful signals benefit receivers, and false signals harm receivers. However, it should be apparent that they do not always coincide in the human case: truthful signals may hurt or cause harm to receivers, and false signals may benefit the receiver.

These two conceptions, or 'dimensions', of honesty then give rise to four possible communicative options one might choose: "co-operative honesty, harmful honesty, co-operative lying and harmful lying" (Dor, 2017, p. 45). While antisocial, exploitative lies – lies by which the communicator profits at the expense of the receiver, i.e. harmful lies – constitute the most intuitive, salient conception of a lie, they are not at all the most prevalent kind of lie. Moreover, Meibauer (2018) notes that prosocial lying is connected to politeness. The notion of politeness, in turn, can be connected with the notion of benefits and costs:

> In antisocial (mendacious) lying, only the speaker profits from lying. In prosocial lying, lying is either altruistic (only the hearer profits from the speaker's lie) or polite ("Pareto-white," as Erat & Gneezy 2012 call it) in the sense that both speaker and hearer profit from the lie. (Meibauer, 2018, p. 371)

Returning to Grice's cooperative principle, note that the above analysis only considers Grice's maxim of Quality ('Try to make your contribution one that is true'). If we extend the above account to the whole cooperative principle – in other words, if we also incorporate the maxims of Quantity, Relation and Manner – we end up with two 'dimensions' of cooperativeness that we might term *Gricean cooperativeness* and *dispositional cooperativeness*. One is then cooperative or noncooperative in a Gricean way depending on whether or not they adhere to Grice's cooperative principle. On the other hand, one is helpful or harmful in their disposition depending on whether their act of communication benefits the receiver or harms them. As with Dor's (2017) account of honesty, these two dimensions then combine to give us four communicative options; see Table 2.1 for an overview with examples for each option.

To fully appreciate the cooperative function of communication, let us now consider what makes cooperation itself evolutionarily beneficial. Moreover, in order to complete the causal chain, we will have a look at how cooperation could have evolved and the role that communication plays in it. We will do so by drawing extensively from Michael Tomasello's *Why We Cooperate* (2009).

		Gricean cooperativeness	
		cooperative	**noncooperative**
dispositional	**helpful**	informative true statement	white lie
cooperativeness	**harmful**	harsh true statement	malevolent lie

Table 2.1: Examples for each of the four communicative options from the two dimensions of cooperation.

dimensions-of-cooperation}

2.4.1 Human cooperation and its evolution

{sec:comm:cooperation}

Let me start off with a brief terminological aside: although colloquially the terms 'cooperation' and 'collaboration' are more or less synonymous, Tomasello does not use these terms interchangeably. He defines collaboration as working together for mutual benefit (2009, p. xvii). Implicitly, he takes cooperation to be an overarching term which also encompasses for example altruism, in which one individual sacrifices something to help another individual. For the remainder of this thesis, I will adhere to these terminological conventions.

Tomasello argues that somewhere along the evolutionary timeline, humans must have been "put under some kind of selective pressure to collaborate in their gathering of food–they become obligate collaborators–in a way that their closest primate relatives were not" (Tomasello, 2009, p. 75). Notably, what exactly this selective pressure is, is a missing link in this story. In general, evolution may select for sociality in animals because living together in a social group protects the group's members against predation: it is easier to defend oneself in the context of a group. The group however also has its disadvantages when it comes to foraging for food, since group's members have to compete with each other for food. This is especially the case when the source of food is 'clumped', such as in a prey animal, rather than dispersed, such as in a plain of grass. The clumped source of food raises the issue of how to share the food amongst the members of the social group. Tomasello enumerates a number of different hypotheses to explain how humans could have broken out of what he calls "the great-ape pattern of strong competition for food, low tolerance for food sharing, and no food offering at all" (Tomasello, 2009, p. 83); in other words, how humans could have evolved to be more tolerant and trusting, and less competitive about food. Firstly, as due to a certain selective pressure it became necessary for humans to forage collaboratively, it would have been evolutionarily advantageous to be more tolerant and less competitive. Secondly, Tomasello notes it could be the case that humans went through a process of self-domestication, by which aggressive, predatory or greedy individuals were eliminated from the group; see also Benítez-Burraco et al. (2021) and Hare (2017). Thirdly, the evolution of tolerance and trust could be related to what is called *cooperative breeding*, also known as *alloparenting*. In cooperative breeding, the responsibility of child-rearing falls on more individuals than just the mother of the child; these individuals help by providing food for the child and engaging in other acts of childcare. This cooperative breeding may have selected for pro-social

skills and motivations; see Hrdy (2009) for an elaboration of this argument.

Tolerance and trust then constitute a foundation upon which coordination and communication could be 'built', so to speak: they provide an environment in which more elaborate collaboration could evolve. In Tomasello's words,

> there had to be some initial emergence of tolerance and trust (...) to put a population of our ancestors in a position where selection for sophisticated collaborative skills was viable (p. 77)

In order to then arrive at the full picture of human cooperative activity, the final step to consider is that of social norms and institutions. In this part of the story, there is also a missing link, concerning how mutual expectations between individuals arise and eventually become norms; Tomasello describes this as "one of the most fundamental questions in all of the social sciences" (p. 89). Norms may be defined as "socially agreed-upon and mutually known expectations bearing social force, monitored and enforced by third parties" (Tomasello, 2009, p. 87). Norms receive their force not only from the threat of punishment by others if the norm is violated, but also from a kind of 'social rationality' in collaborative activity, where individuals recognize their dependence on each other in reaching their joint goal. Just as it would be individually irrational to act in a way that thwarts your own goal, it would be socially irrational to act in a way that thwarts your joint goal.

Let us now summarize the evolutionary timeline of human cooperation according to Michael Tomasello. At some point in time, for reasons as of yet unknown to us, foraging for food collaboratively rather than individualistically became beneficial – perhaps even necessary – for humans. As a result of this, some degree of tolerance and trust must have emerged between these collaborating individuals. In the process of adapting to this collaborative foraging, humans evolved certain skills and motivations specifically for cooperation – for example, abilities for establishing joint goals and role divisions for joint activity. This kind of collaborative activity then constituted the breeding ground for human cooperative communication. Joint goals and role divisions later evolved into the superindividual norms, rights and responsibilities that we see within our social institutions today.

As a brief aside: it has been argued that communication is not necessary nor sufficient for the coordination of activities. Goldstone et al. (2024) propose a framework of five features characterizing the specialization of roles in group activities; communication is only one of these five features. This is corroborated by experiments they review in which people "spontaneously differentiate themselves into stable roles" (p. 264) in group activity without communicating with each other. However, the authors note that communication does play a very central role in coordinating group activities, stating that "direct communication of plans is often the single most potent tool of collective coordination" (Goldstone et al., 2024, p. 276). See also Vorobeychik et al. (2017) for a discussion of communication and coordination.

Now, armed with the ins and outs of human cooperation, and an inkling of how communication relates to it, we turn our attention to a crucial aspect

in understanding the evolution of human communication: the problem of the stability of communication. This problem in a way constitutes the starting point for the argumentative theory of reasoning, as we will see in Section 3.1.

2.4.2 The stability of communication

{sec:S-P08}

If communication between individuals of a species persists throughout evolution, we may speak of it as stable. The stability of communication is considered by some to be the 'defining problem' of animal signaling research (Scott-Phillips, 2008). It is not a trivial problem by any means: the stability of a communication system is threatened by evolutionary pressures on the communicator to 'defect', as it were. As Scott-Phillips (2008) describes it,

> If one can gain through the use of an unreliable signal then we should expect natural selection to favour such behaviour. Consequently, signals will cease to be of value, since receivers have no guarantee of their reliability. This will, in turn, produce listeners who do not attend to signals, and the system will thus collapse in an evolutionary retelling of Aesop's fable of the boy who cried wolf. (p. 275)

In the context of human communication, this problem may be interpreted as follows: if it can be advantageous for me to deceive you, then it would evolutionarily speaking make sense for me to do so; yet, it would then evolutionary speaking make sense for you to stop listening to me, and as a consequence our system of communication would collapse.

There have been a number of attempts at explaining the stability of animal communication in general. One influential attempt is the *handicap principle* (Zahavi, 1975; Zahavi and Zahavi, 1999), which might be best understood through the paradigmatic example of the peacock's tail. This tail is like a handicap for the peacock: not only does it take a lot of resources to grow the tail and carry it around, it also leaves the bird more vulnerable to predation because it is less agile with a large unwieldy tail. At the same time, a large tail signals to peahens that the peacock is fit enough to be able to incur these costs, and thus the peacock has a sexual advantage. The handicap principle then describes this process of communication, by which the signaler incurs costs (i.e., a handicap) for signaling, which thus guarantees the reliability of the signal.

While influential, the handicap principle has been argued to fall short of explaining the stability of communication (Penn and Számadó, 2020; Scott-Phillips, 2008). One reason for this is that reliable signals are not always costly to produce. For example, Harris sparrows signal their social status to other sparrows by the amount of black feathering in their plumage, and this signal is not costly to produce (Lachmann et al., 2001, p. 13191). Moreover, human linguistic communication is considered to be generally[2] a low-cost endeavor: "Although the capacity for speech may itself be costly, once language has been

[2]See Lachmann et al. (2001, p. 13193) for a discussion of when and why humans may use costly forms of communication.

acquired the production costs vary little among alternative signals" (Lachmann et al., 2001, p. 13192). Thus, it remains to be shown how communication can be stable if signals are cheap.

An alternative explanation of the reliability of animal communication is the principle of *deterrence*, where *un*reliable signals are costly to produce, and consequently signalers are deterred from producing unreliable signals (Scott-Phillips, 2008). There are a number of ways in which producing unreliable signals may be costly to the signaler. Firstly, this is the case in a coordination game, where the signaler and receiver share some common interest with regard to the outcome of the interaction (see Maynard Smith, 1994). Secondly, if two individuals have repeated interactions, it may also be costly in the long term to produce unreliable signals, because it may hinder cooperation in the future. Thirdly, producing unreliable signals may be costly to the signaler if false signals are punished by the receiver. In the case of the Harris sparrow for example, subordinate sparrows that are dyed black (and are thus signalling falsely) are socially persecuted by other sparrows (S. Rohwer and F. C. Rohwer, 1978).

This 'logic of deterrents' can be applied to the case of human communication to account for its stability, if the following demands are met:

> Sufficient conditions for cost-free signalling in which reliability is ensured through deterrents are that signals be verified with relative ease (if they are not verifiable then individuals will not know who is and who is not worthy of future attention) and that costs be incurred when unreliable signalling is revealed. (Scott-Phillips, 2008, p. 279)

In other words, if unreliable signals are 'caught' relatively easily, and unreliable signalers incur costs for their unreliability, then the reliability of communication is secured through deterrents. Scott-Phillips (2008) argues that these sufficient conditions are met in the case of human communication, since people may refrain from interacting with unreliable individuals in the future, which can be very costly for a social species such as humans. Notably however, he does not explicate how exactly the first sufficient condition is met in the case of human communication. We will return to this in Section 4.2, as it turns out that the ease with which unreliable signals are caught, is a contentious point of the ATR.

The problem of the stability of communication hinges on the assumption that it is advantageous to deceive others through the use of unreliable signals. Let us explore this further by turning our discussion to deception and lying.

2.4.3 Deception and lying

{sec:deception}

Let us start off this section by clarifying and examining some terminology. In our discussion of the stability of communication, we have so far been talking about 'reliability' rather than 'honesty'. This is because 'reliability' is a more neutral term than 'honesty', in the sense that it refrains from ascribing intentions and meanings to individuals; and especially in the case of animal communication, we should steer clear of ascribing intentions and meanings to individuals (Scott-Phillips, 2008). This ascription of intentions and meanings is

not problematic when discussing human communication however; moreover, I find that talking about 'reliable communicators' blurs an important distinction between honest, benevolent, and competent communicators. We return to this distinction in Section 3.3.

Deception and lying and may be defined in a number of different ways. First off, deception may be defined as "deliberately leading someone into a false belief" (Meibauer, 2018, p. 358). Lying, on the other hand, might not be as easily defined. Consider the following 'standard' definition of lying (Meibauer, 2018), due to Williams (2002):

> an assertion, the content of which the speaker believes to be false, which is made with the intention to deceive the hearer with respect to that content (Williams, 2002, p. 96)

There is, however, a case to be made for excluding 'intention to deceive' from this definition. In the case of bald-faced lying, where the liar knows he will not be believed, or in cases where people are forced to lie (in totalitarian states, for example), lying occurs without intention to deceive (Saul, 2012, §1.5). For the purposes of this thesis and the general nature of its discussion however, the exclusion of these edge cases is not problematic. Thus, we will proceed using Williams's (2002) definition and assume that lying is by definition done with deceitful intent.

Let us now connect lying with the problem of the stability of communication. In this narrower context, let us refer to this problem more concisely as the 'paradox of honest signaling' (following Dor, 2017). Dor (2017) argues that the foretold collapse of communication due to unreliability of the speaker does not hinge on whether or not the speaker is truthful, but whether her *intention* is benevolent. In other words, it matters to the story not that receivers evaluate the truthfulness of incoming information, but rather the *intention* of the sender. Listeners care whether speakers intend to be harmful, not whether or not they are truthful, Dor argues. Moreover, he notes that the paradox of honest signaling mostly applies to situations in which interests between interlocutors conflict; however, these situations might not be the most pertinent or prevalent kind of communicative situation. He argues that, at the point in evolutionary time when language emerged, humans were already crucially dependent on cooperation and coordinated action, and thus their interests overlapped more often than not.

There is thus an interesting interaction to observe between the paradox of honest signaling and the utility of communication. As we will see in Chapter 3, the argumentative reasoning mostly focuses on communication as the transmission of information between individuals, in the form of testimony and argumentation. However, within the paradox of honest signaling this transmission of information is not the only relevant use of communication. Communication is also used for cooperation, and in that situation lying is not really an issue, as argued by Dor (2017):

> Language is extremely useful in the coordination of collective work, collective defense and so on, where it is used not just for the exchange of in-

> formation but also for collective planning, division of labor, ordering and requesting, where lying as such does not seem to play a major role. (Dor, 2017, p. 51)

Therefore, due to this dual role of communication (transmitting information on the one hand, and facilitating cooperation on the other), the stability of communication is not threatened by lying:

> Even in the very unlikely doomsday scenario, then, where all the members of a community lie to each other in their factual statements, and eventually refrain from sharing information with each other, there is no reason to assume that they would stop using language for all these other purposes, especially where their survival, whether they like it or not, depends on collective action. (Dor, 2017, p. 52)

We return to these considerations in Chapter 4.

Lastly, we will have a brief look at persuasion, since it naturally plays a considerable role in Mercier & Sperber's argumentative theory of reasoning. Brinol and Petty (2009) broadly define persuasion as "any procedure with the potential to change someone's mind" (p. 50), whether that be changing someone's emotional state, beliefs, behaviors or attitudes. They describe persuasion as "the most frequent and ultimately efficient approach to social influence" (pp. 49–50). Put crudely, persuasion is a tool for getting what you want, and it serves this end better than the alternatives of using force, threats or violence.[3] From this observation, the conclusion emerges that persuading someone is beneficial to an individual exactly to the extent that the corresponding gain in social influence is beneficial to the individual.

2.5 Conclusion

This chapter saw us examining the evolution of human communication by studying each of the methodological questions raised in Section 1.6. We sketched an overview of different conceptions of communication, had a look at how non-human animals and human children communicate, and we considered the utility of communication at length. In doing so, we outlined different dimensions of cooperation and discussed the evolution of human cooperation. Moreover, we introduced the problem of the stability of communication, and discussed lying and deception.

From the analysis in this chapter, the view arises that the uniquely human way of communicating evolved as a result of humans' growing dependence on cooperation for survival. We will return to this view in Chapter 4; but first, let us have a detailed look at the argumentative theory of reasoning.

[3] In this, one may see a parallel with Seyfarth and Cheney's (2003) conception of aggressive communication (intimidation) as a low-risk alternative to fighting.

3 | The argumentative theory of reasoning

{ch:atr}

The argumentative theory of reasoning is Hugo Mercier and Dan Sperber's influential, but not uncontroversial, account of the function of reasoning from an evolutionary perspective. They introduced and coined the theory in a 2011 paper, a culmination of more than a decade's worth of research (Sperber, 2001, Sperber et al., 2010, Mercier and Sperber, 2009). Briefly, the argumentative theory of reasoning states that the main function of reasoning is to produce arguments and evaluate arguments of others, for the purpose of stabilizing communication.

In this chapter, I will expound the argumentative theory of reasoning by discussing at length each of the precursory papers leading up to Mercier and Sperber (2011), in chronological order. In Section 3.1, we will discuss Sperber's 2001 contribution *An evolutionary perspective on testimony and argumentation*. In Section 3.2, we tackle Mercier and Sperber's dual system theory, put forward in the 2009 paper *Intuitive and reflective inferences*. Next, Section 3.3 sees us discussing *Epistemic vigilance*, a 2010 paper by Sperber, Mercier, and colleagues. By this point, much of the ATR is already laid out. Finally, in Section 3.4, we will summarize the ATR from Mercier and Sperber (2011), and briefly discuss the empirical predictions of the theory.

3.1 Sperber on the evolution of testimony and argumentation

{sec:Sperber01}

In a 2001 paper, Dan Sperber analyzes testimony and argumentation from an evolutionary perspective. In doing so, he provides important groundwork for his later work with Mercier (and others) on the relation between reasoning, argumentation and the stability of communication.

Testimony and argumentation are two concepts central to human communication. Sperber borrows his definitions for these concepts from epistemologist Alvin Goldman, who defines testimony as "the transmission of observed (or allegedly observed) information from one person to others" (Sperber, 2001, p. 401) and argumentation as "the defense of some conclusion by appeal to a

set of premises that provide support for it" (ibid.). Sperber puts these two concepts in an evolutionary perspective, and discusses in particular how they have figured in stabilizing communication over the course of evolutionary history.

A tempting way to look at communication is as a kind of 'cognition by proxy': through communication, one organism may access information another organism has obtained from its own perception or inference. For instance, if one vervet monkey expels an alarm call upon observing a leopard in the distance (Seyfarth et al., 1980), its conspecifics can through this act of communication benefit from the information derived from the alarmed monkey's perception of the leopard in the distance. However, Sperber argues that in the case of human communication, testimony does not amount to cognition by proxy. This is because in humans, testimony has different effects than direct perception does. Suppose I observe a leopard in the distance and inform you of this. Upon receiving this testimony, you are in a different cognitive state than you would be if you had perceived the leopard yourself, Sperber argues. Moreover, in human communication, he maintains that interpretation and acceptance of utterances are two separate processes: recognizing what a speaker meant by their utterance is not the same as accepting it as true.[1]

The classical account of animal communication by Dawkins and Krebs (1978) focuses only on the side of the communicator in the story, maintaining that the function of communication is to manipulate others. Sperber rejects this classical approach, arguing that the interests of the sender cannot be the only driving force in the evolution of communication. He outlines a similar line of argumentation as we have seen in Section 2.4.2, arguing that for communication to have stabilized and continued to be stable between senders and receivers, both parties must have benefited from it. In other – game-theoretic – terms, communication must (at least in the long run) be a positive-sum game, where both senders and receivers gain from the interaction.

In the case of receiving testimony from others, the receiver gains from testimony "only to the extent that it is a source of genuine (...) information" (p. 404). On the other side, a sender stands to gain from the production of testimony (and from communication in general) because

> it allows them to have desirable effects on the receivers' attitudes and behavior. By communicating, one can cause others to do what one wants them to do and to take specific attitudes to people, objects, and so on (Sperber, 2001, p. 404)

Sperber later elaborates on this by stating that getting others to accept your communicated message is not intrinsically beneficial. Rather, it is indirectly beneficial, through bringing about these 'desirable effects' in others, as a way

[1] This may very well be the case philosophically or epistemologically speaking, but psychologically speaking, they may be more intertwined than Sperber implies. In Sperber et al. (2010, §3), he and his colleagues elaborate more on his stance, making even stronger claims about how comprehension always precedes the acceptance (or rejection) of a claim. Although this is intuitively plausible, see also Gilbert (1991) advocating that for someone to comprehend an utterance, they must (at least temporarily) accept it.

of *cognitive manipulation*. Sperber notes that it is exactly this self-interest of the sender that renders this 'cognition by proxy' view as inapplicable to human communication. Moreover, he concludes from his observations that

> the function of communication presents itself differently for communicator and audience (p. 411)

We return to this conclusion to criticize it in Section 4.1.

Sperber goes on to cast his observations in game-theoretic terms by sketching out a payoff matrix for a one-off communicative event; see Table 3.1.

		addressee	
		trusting	**distrusting**
communicator	**truthful**	gain/gain	loss/no gain
	untruthful	gain/loss	loss/no gain

{tab:matrix}

Table 3.1: Payoff matrix for a one-off communicative event

In sketching out this scenario, he considers that senders may be truthful or untruthful, and receivers may be trusting or distrusting. According to Sperber, the sender's gain amounts to whether they have the 'desired' effect on the receiver; therefore, the sender gains from the interaction if the receiver is trusting (since this means the sender's message is accepted), and loses from the interaction if the receiver is distrusting. The payoff of this event for the sender is thus independent of the truthfulness of the sender. On the side of the receiver, their payoff *is* dependent on the truthfulness of the sender: the receiver gains if they accept a truthful message, loses if they accept an untruthful message, and incurs no gain nor loss if they are distrusting and thus don't accept a message (truthful or not).

Sperber notes that the optimal strategy for a game like this varies with the circumstances for both players: it is not beneficial to be always truthful, nor always untruthful; nor is it beneficial to be always trusting, nor always distrusting. In other words, there is no one stable solution to this game. This is especially the case once we move away from this simple one-off communicative event to an iterated game of communication, where not only short-term payoffs but also long-term payoffs determine the optimal strategy. Therefore, it is in the receiver's interest to calibrate their trust towards senders as accurately as possible; in fact, Sperber argues, this trust calibration is necessary to account for the stability of communication.

Unlike non-human animals, humans can not only communicate facts through testimony; they also have *argumentation* at their disposal. Senders may provide receivers with reasons to accept their testimony; the receiver may then evaluate these reasons and accept or reject the testimony, independent of their trust in the sender. Sperber notes that reasoning may be used individually for reflection, or socially for communication in dialogical argumentation. He argues that, although classically the former has been viewed as the function of reasoning (cf. Novaes (2020)), this is implausible from an evolutionary point of view.

He maintains that domain-general reasoning abilities are cognitively costly and slow, and therefore could not have evolved for the purpose of producing knowledge, since more specific mechanisms would better suit this function. Instead, the function of reasoning is communicative rather than individual:

> [Reasoning] is an evaluation and persuasion mechanism, not, or at least not directly, a knowledge-production mechanism. (Sperber, 2001, p. 409)

Next, Sperber sketches out the steps in what he calls the 'evaluation-persuasion arms race', i.e., the chain of evolutionary adaptations that has resulted in our mechanisms for argument production and evaluation. He argues that the first step in this 'arms race' was for the receiver to develop *coherence checking*. Coherence checking involves attending to both the internal coherence of the communicated message, and the external coherence with what the receiver already believes. Coherence checking, Sperber argues, is a useful defense against the risks of being deceived by the sender, because lies and other false claims are often externally or internally incoherent. The second step in the arms race was then for the sender to anticipate this coherence-checking by overtly displaying the coherence of their message to their receiver, which requires argumentative form; thus, testimony becomes argument. The next steps in the arms race were then on the side of the receiver to develop skills for examining these displays of coherence (i.e., arguments), and on the side of the sender to 'improve their argumentative skills'. Mercier and Sperber nicely capture these next steps in the arms race in their 2011 paper, stating that

> receivers' coherence checking creates selective pressure for communicators' coherence displays in the form of arguments, which in turn creates selective pressure for adequate evaluation of arguments on the part of receivers (p. 96)

3.2 Mercier & Sperber on intuitive vs. reflective inference

{sec:MS09}

Let us move on to the next stop along the path to the argumentative theory of reasoning. In the 2009 paper *Intuitive and reflective inferences*, Mercier and Sperber propose their own dual system theory of reasoning, in a similar vein to existing dual process theories of reasoning (Evans, 2003; Evans and Stanovich, 2013; Kahneman, 2011; Sloman, 1996). They introduce their distinction between intuitive and reflective inferences as part of a massive-modularist framework, which maintains that the mind consists of a number of cognitive modules specialized to specific domains. This view is certainly not an uncontroversial one (cf. Dutilh Novaes (2018, §3.1)); I consider a detailed discussion of the massively-modularist view of the mind out of scope of this thesis.

Before we consider the contributions by Mercier and Sperber (2009) however, let me make a slightly anachronistic pivot. In order to clarify some terminology that is not adequately discussed in Mercier and Sperber (2009), we

Change this word

briefly turn to a discussion of intuitive inference and argument in Mercier and Sperber (2011, §1.1). There, Mercier and Sperber posit that the processes that are executed by inferential modules are unconscious: though one might be aware of the output of such a process – its conclusion – one is unaware of the process itself. In other words, "All inferences carried out by inferential mechanisms are in this sense *intuitive*" (Mercier and Sperber, 2011, p. 58). Importantly, they then highlight a distinction between inferences and arguments. Inferences are processes: they take as input a representation, and they output a representation. Arguments on the other hand are representations themselves, resulting from inference. They are "the output of an intuitive inferential mechanism"; in particular, they are "representations of relationships between premises and conclusions" (Mercier and Sperber, 2011, p. 58). Both inferences and arguments have conclusions, but there is an ontological dissimilarity between these conclusions. The conclusion of an inference is its output. Characteristically, the output of an inference is justified by the input of the inference; thus, we call this output a conclusion. The conclusion of an argument, on the other hand, is part of the representation itself, i.e. it is part of the argument.

Pivoting back to Mercier and Sperber's 2009 discussion of intuitive and reflective inferences, the authors argue that one of the many modules of the mind is the argumentation module, which "provides us with reasons to accept conclusions" (p. 155). The module takes as input a claim, and potentially other information relevant for evaluating this claim, and it produces as its output reasons for accepting or rejecting the claim. The authors note that the direct output of any inferential module is *intuitive*, in the sense that we accept the module's output without consciously attending to the reasons for this acceptance. This is then also the case for the argumentation module, as it is an inferential module like any other. The direct ('intuitive') output of the argumentation module is then "the representation of an argument-conclusion relationship" (ibid.). The difference then between intuitively accepting a claim and accepting it because of explicit reasons, is that in the latter case one engages in "disembedding a conclusion from the argument that justifies it" (ibid.). This disembedded conclusion from the direct output then constitutes the indirect output of the argumentation module.

From this view on the modularity of the mind, and the argumentation module in particular, Mercier and Sperber then develop a dualistic approach to inference. Their account distinguishes between intuitive inferences on the one hand, and reflective inferences on the other. Reflective inferences are then what Mercier and Sperber refer to as reasoning, or 'reasoning proper'.

Mercier and Sperber emphasize that their dual systems approach is different from the classical distinction between the fast and frugal thinking of system 1 and the slow and analytical thinking of system 2 (cf. Kahneman (2011)). They maintain that system 1 and system 2 operate at the same level[2], whereas intuitive and reflective inference do not: intuitive inferences are carried out by any module, but reflective inferences result specifically from the argumenta-

[2]It is unclear what they mean by this comment though; see Mercier and Sperber (2009, p. 156).

tion module – more precisely, reflective inferences are an indirect output of the argumentation module (Mercier and Sperber, 2009, p. 156).

Moving beyond their dual system approach, Mercier and Sperber go on to provide yet more foundation to their up-and-coming argumentative theory of reasoning by discussing the function of reflective inference – in other words, the function of reasoning. Before doing so, they remark that the function of intuitive inference is less controversial than the function of reflective inference (but they do not explicate what this function is). With regards to the function of reflective inference, they first discuss three 'classical' views on the function of system 2 reasoning. The first view maintains that system 2 represses system 1's impulses seeking immediate gratification, in order to obtain delayed gratification (Sloman, 1996). Mercier and Sperber argue that this cannot be the function of reasoning, since non-human animals also possess the ability to delay gratification, but not the ability to reason[3]. Moreover, they argue that empirical case studies discussed in Damasio (1994) suggest that abilities for delayed gratification are dissociated from abilities for reasoning. The second view on the function of reasoning maintains that system 2 reasoning enables us to better deal with novelty (Evans and Over, 1996). Mercier and Sperber argue that it is implausible that this is the function of reasoning, since there are other features of human cognition that better explain and support this ability. Moreover, they state that reasoning cannot be said to play a central role in memory and imagination.

These two views on the function of reasoning have in common that they posit that system 2 'compensates' for the shortcomings of system 1. This idea culminates in the third view on the function of reasoning that Mercier and Sperber discuss. This is the view that the function of reasoning is to enhance individual cognition, or in a stronger version, this view concerns the Cartesian conception of reasoning as '*the* road to knowledge'. They argue that this is evolutionarily implausible due to a cost-benefit issue. Reasoning is cognitively costly, and intuitive inference is not so unreliable that using reflective inference instead would be, on the whole, advantageous.

In the remainder of the paper, the authors discuss predictions stemming from their ascribed function of reasoning, and empirical evidence corroborating these predictions. We will return to these and more empirical predictions of the ATR in Section 3.4, since Mercier and Sperber's 2011 paper goes into much more detail on this.

3.3 Sperber and colleagues on epistemic vigilance

{sec:Sperber10}

Next up is the concept of *epistemic vigilance*, which was introduced by Sperber, Mercier and colleagues in a seminal 2010 paper. This concept constitutes

[3]This latter claim is left implicit by Mercier and Sperber in this paper. In Mercier and Sperber (2011, p. 57), they explicate their conviction that non-human animals cannot reason; cf. Hume (1739/1978).

a cornerstone of the argumentative theory of reasoning. We will return to the concept of epistemic vigilance to evaluate it critically in Section 4.2.

Sperber et al. (2010) start off by emphasizing that humans are dependent on communication. They argue that this dependence leaves humans vulnerable to being deceived by others, stating that misinformation or deception may "reduce, cancel, or even reverse" the gains that communication can bring to the addressee (p. 360). Consequently, the information that an addressee receives from a communicator is only advantageous to her to the extent that the information is genuine. Sperber and colleagues thus conclude that for this purpose, humans have evolved a suite of cognitive mechanisms for *epistemic vigilance*. Moreover, this suite of mechanisms must have evolved alongside, and is used in tandem with, abilities for ostensive-inferential communication[4], because they work in tandem to facilitate trust calibration on the side of the receiver.

First, Sperber and colleagues discuss work from both classical and contemporary epistemology and experimental psychology on trust; specifically, they consider different views on the question of whether humans are 'per default' trusting or vigilant. They conclude that one can acknowledge the importance of epistemic trust in communication while simultaneously acknowledging the importance of epistemic vigilance, they co-exist because

> Vigilance (unlike distrust) is not the opposite of trust; it is the opposite of blind trust (p. 363)

Indeed, not only can epistemic trust co-exist with epistemic vigilance, it is buttressed by it, they argue, concluding that

> We could not be mutually trustful *unless* we were mutually vigilant. (p. 364)

We return to this claim in particular in Section 4.2.

Next, the authors move on to discuss comprehension and acceptance of utterances in communication, and how these relate to epistemic vigilance and trust. They argue that a communicative act does not only trigger comprehension in the addressee, but it also triggers epistemic vigilance alongside it. If epistemic vigilance then "does not come up with reasons to doubt" (p. 369), comprehension leads to acceptance. They go on to argue that comprehension of an utterance is not "guided by a presumption of truth", as other theorists state, but rather by an "expectation of relevance" (p. 367); see Sperber and Wilson (1986). This expectation of relevance requires a 'stance of trust' of the addressee regarding the speaker (this relates to the Gricean cooperativeness we discussed in Section 2.4). This stance of trust of the addressee is "tentative and labile" (p. 368), and epistemic vigilance is (as mentioned) active alongside this stance of trust.

[4]Slightly confusingly, Sperber et al. call the ostensive-inferential communication that we saw in Section 2.1 'overt intentional communication' in this paper. However, they cite Sperber & Wilson's *Relevance Theory* (1986), which uses the term 'ostensive-inferential communication'. So ultimately, they are talking about the same thing.

To further explicate epistemic vigilance as a concept, Sperber and colleagues outline a distinction between vigilance towards the *source* of a message (the 'who'), and vigilance towards the *content* of the message (the 'what'). As for vigilance towards the source, they note that the reliability of a source depends on two factors: a reliable source must be competent, and a reliable source must be benevolent. Moreover (and importantly), a receiver's vigilance towards the sender as a source of information – in other words, the sender's perceived trustworthiness – is dependent on the context: it varies per topic and per situation. Because of this, it is important for a receiver to accurately calibrate her trust in the sender, depending on the context. The authors go on to discuss empirical evidence that corroborates that trust and trust calibration is indeed important to us. Moreover, on the other side of the coin, they note that deceiving people can be quite beneficial, because liars are not easily caught: experiments from deception detection research show that people are not good at detecting lies based on non-verbal behavioral cues (see e.g. Vrij (2000)). They end this discussion by noting that more empirical research is needed about how people calibrate their trust in everyday communication, outlining some desiderata for this research.

Moving on now to vigilance towards the *content* of a message, Sperber and colleagues restate that comprehension and epistemic vigilance are two processes that are intertwined to some extent. Specifically, they note that one mechanism of comprehension, namely the search for relevance, provides a basis for an "imperfect but cost-effective epistemic assessment" (p. 374). They discuss belief revision and the role that coherence checking plays in it. We already saw Sperber's (2001) discussion of coherence checking; Sperber and colleagues now declare coherence checking a mechanism for epistemic vigilance. They note that coherence checking "takes advantage of the limited background information activated by the comprehension process itself" (p. 375). They argue that the search for relevance "automatically involves the making of inferences which may turn up inconsistencies or incoherences relevant to epistemic assessment" (p. 376).

To summarize, according to Sperber, Mercier and colleagues, humans have developed a suite of mechanisms for epistemic vigilance, filtering incoming information in order to avoid being deceived by others. A communicative act triggers both comprehension and epistemic vigilance, and the epistemic assessment of the communicative act draws upon some of the inferential steps that are carried out in the search for relevance, which makes the assessment relatively cost-effective. Epistemic vigilance can be directed towards the source of a message or towards the content of the message. This amounts to the calibration of trust and coherence checking, respectively.

3.4 Mercier & Sperber's argumentative theory of reasoning

{sec:MS11}

Now, the key features and aspects of Mercier and Sperber's theory were already laid out either in detail or in a rudimentary form in the papers we have discussed in the previous sections. In this section, I will first summarize the full argumentative theory of reasoning, and then consider the theory's empirical predictions and the corroborating evidence Mercier and Sperber (2011) bring to the table.

3.4.1 The ATR summarized

In order to get a good overview of the ATR, it will be illuminating to consider the theory in a schematic way, in something resembling argument form. (The following is paraphrased primarily from the exposition of the ATR in Mercier and Sperber (2011, p. 60).)

(1) For their survival, humans are dependent on cooperation with other humans, and communication is crucial for this: "Communication plays an obvious role in human cooperation both in the setting of common goals and in the allocation of duties and rights" (Mercier and Sperber, 2011, p. 60).

(2) For communication between humans to have stabilized over the course of evolutionary history like it has, it must have been advantageous to both the senders and the receivers of messages. If it were not, the practice would have collapsed over time (Sperber, 2001).

(3) It is advantageous for senders to be dishonest – "communicators commonly have an interest in deceiving" (Mercier and Sperber, 2009, p. 160). Frequent deception threatens the stability of communication, because it renders communication disadvantageous to the receiver (Sperber, 2001).

(4) To protect themselves against deception, receivers need to (and have therefore evolved the means to) exercise *epistemic vigilance* in order to filter incoming information (Sperber et al., 2010).

(5) In order to then get their message across to a vigilant receiver, a sender may demonstrate the coherence of her claims by offering an argument as a reason to accept her claim.

(6) This demonstration of coherence – i.e., this argument – is produced by reasoning, and the evaluation of the argument by the receiver is also facilitated by reasoning.

(7) Thus, the function of reasoning is the production of arguments (by the sender) and the evaluation of arguments (by the receiver). In other words,

> reasoning has emerged and persisted throughout evolutionary history precisely because it enables the production and evaluation of arguments. Argumentation plays a critical role in ensuring the stability of communication, which ultimately contributes to humans' survival.

3.4.2 Empirical predictions and evidence

Mercier and Sperber (2011) point out that evolutionary hypotheses are at risk of coming across as "just so stories" if they are not buttressed by empirical evidence. They argue that

> To establish that reasoning has a given function, we should be able at least to identify signature effects of that function in the very way reasoning works. (p. 60)

Consequently, they outline four predictions that follow from the ATR and they present empirical evidence that corroborates these predictions.

The first prediction of the ATR is that reasoning is best adapted to perform tasks in argumentation. In other words, reasoning is good at producing and evaluating arguments, and it works best in an argumentative context, e.g. in group discussions. Many classical findings from the psychology of reasoning, such as the Wason selection task (Wason, 1968), conclude that people are poor logical reasoners. Mercier and Sperber note however that people's performance on this kind of task improves if it is moved from a nonargumentative to an argumentative setting. They cite evidence that people are generally good at spotting fallacies in others' arguments , and that they are skilled at recognizing the structure of arguments . Moreover, the authors discuss findings on group reasoning tasks, in which participants are first tasked with solving problems individually, then in a small group, and lastly individually again. On for example the Wason selection task, participants' performance improved dramatically in a group setting (Moshman and Geil, 1998) .

Add reference

Add reference

The second prediction is that reasoning shows a confirmation bias.

Doing what? Give one example

Thirdly, the ATR predicts that solitary reasoning anticipates argument, in a form of *motivated reasoning*.

Add explanation and references

Lastly, the ATR predicts that reasoning in decision-making guides people not to the optimal decision, but to the decision that is most easily justified.

Add explanation and references

3.5 Conclusion

Add explanation and references

Over the past decades, Hugo Mercier and Dan Sperber have dedicated an impressive amount of work to carving out their thesis on reasoning's function of empowering and improving communication. Now that we have dissected the ATR, the time is upon us to take a critical look at a number of the theory's component parts. Because, however appealing the argumentative theory of reasoning sounds, it becomes less appealing the more you look at it.

4 | The ATR closely inspected

{ch:scrutiny}

Although Mercier and Sperber's story about the evolution and function of reasoning as it relates to communication is prima facie quite convincing, it seems to leave a number of details underexposed. In this chapter, I will scrutinize these details and highlight areas for improvement.

In Section 4.1, I will critically evaluate Mercier and Sperber's purported function of communication by stacking it against the picture painted in Section 2.4. Section 4.2 will see us discussing epistemic vigilance at length by considering a critical response paper to Sperber et al. (2010), and Sperber's reply to this criticism. Lastly, in Section 4.3 I will detail some metatheoretical frustrations regarding the generally imprudent way in which Mercier and Sperber define and use the concepts they introduce.

4.1 The function of communication: revisited

{sec:comm-func-scrutiny}

As we have seen in Section 2.4, I argue that the function of communication is to facilitate cooperation. Humans are a uniquely cooperative species: we depend on each other for our survival, and communication plays a crucial role in enabling this. On this view, it is critical to consider the perspective of the whole social group, not just the individuals interacting with each other.

The function of communication as it transpires from the work of Mercier and Sperber approaches it from a different angle. As we have seen in Chapter 3, they agree to some extent with my view, as they note the importance of small group cooperation in evolutionary history and the role communication plays in this cooperation (Mercier and Sperber, 2011, p. 60). However, for the most part, their theory divorces itself from this broader context, as they choose to focus only on how communication benefits senders and receivers.

Sperber argues that "the function of communication presents itself differently for communicator and audience" (Sperber, 2001, p. 411). On his view, communication is advantageous for senders to the extent that their communicated message causes "desirable effects on the audience" (Sperber, 2001, p. 406). For this function of communication, it does not matter whether the communicated message is true or not. In fact, Sperber maintains that it is in many cases advantageous for senders to deceive receivers. Sperber et al. (2010) write:

> the major problem posed by communicated information has to do not with the competence of others, but with their interests and their honesty. While the interests of others often overlap with our own, they rarely coincide with ours exactly. In a variety of situations, their interests are best served by misleading or deceiving us. (p. 360)

Moving over to the side of the receiver, Sperber and colleagues argue that communication is advantageous for them because it allows them to gain information.

While both of these claims may indeed be intuitively very plausible, they are underspecified from an evolutionary perspective. What exactly are these 'desirable effects' people might achieve in their audience, and how might these effects improve an individual's fitness? What is exactly the advantage of gaining information? Multiple key steps are missing from the causal chain because they are taken for granted as uncontroversial, weakening the overall account. Moreover, these claims lack connection to a cooperative context.

Because Mercier and Sperber's theory is disassociated from the evolutionary context of human cooperation, the evolutionary story becomes somewhat blurry. Catarina Dutilh Novaes also expresses reservations regarding this aspect of the ATR in her 2018 review of Mercier and Sperber (2017), criticizing Mercier and Sperber's focus on the individual and accompanying dismissal of the group-level perspective (Dutilh Novaes, 2018, §3.3).

This failure to incorporate the essential background of human cooperation does, however, not count as a definitive strike against the theory: I conjecture that the ATR can be integrated with the account of human cooperation we saw in Section 2.4. The ATR and this account are not incompatible per se, but integrating them does most likely require small modifications or additions to the ATR. In particular, the function of communication for senders and receivers needs to be linked to the cooperative function of communication to complete the evolutionary account. Doing this would fortify the ATR: integrating the theory within this cooperative picture would make the story as a whole more plausible, because the (currently questionable) assumptions would then be tethered to a broader context.

4.2 Epistemic vigilance: revisited

{sec:EV-scrutiny}

It appears that the ATR rests on epistemic vigilance having evolved in humans. Therefore, this concept deserves some critical analysis. In this section, I will first discuss the position of the notion of epistemic vigilance within the argumentative theory of reasoning. Then, I will discuss a critical response to Sperber et al. (2010) by Kourken Michaelian (2013), and Dan Sperber's (2013) response to the criticism. Lastly, I will conclude with the consequences of this discussion for the argumentative theory of reasoning.

Before we start, let me first briefly outline Michaelian's paper and Sperber's response to it. In short, Michaelian spells out some of the assumptions that he claims are inherent to Sperber et al.'s argument for epistemic vigilance. He

contrasts these assumptions with empirical findings from deception detection research. In doing so, he concludes that epistemic vigilance does not play a major role in ensuring the stability of communication — rather, he argues, the majority of the burden befalls *speaker honesty*. In the same 2013 issue of *Episteme*, Dan Sperber provides a response to the criticisms of Michaelian. Based on the argument Michaelian put forward, Sperber further explains and sometimes tweaks some of the claims he and his colleagues made in the original 2010 paper. He attributes a considerable portion of Michaelian's criticism to a misunderstanding, which effectively nullifies Michaelian's import of empirical findings.

4.2.1 Epistemic vigilance and the ATR

{sec:epi-vigil-atr}

From our discussions of Sperber et al. (2010) and Mercier and Sperber (2011) in Chapter 3, it transpires that the notion of epistemic vigilance plays some role of importance in the argumentative theory of reasoning. However, in order to assess how gripes with epistemic vigilance as a concept would impact the ATR's credibility, we should consider the exact position of epistemic vigilance within the ATR.

To start off, the paper coining the ATR has the following to say about epistemic vigilance:

> For communication to be stable, it has to benefit both senders and receivers (...) To avoid being victims of misinformation, receivers must therefore exercise some degree of what may be called epistemic vigilance (Sperber et al. 2010). (Mercier and Sperber, 2011, p. 60)

In other words, the stability of communication depends on its benefits for both the sender and the receiver. According to Mercier & Sperber, the benefits for the receiver depend on, or are mediated by, epistemic vigilance.

The evolutionary arms race (recall Section 3.1) is foundational to the argumentative theory of reasoning. Epistemic vigilance is one of the steps in this arms race, having evolved as a defense against misinformation and deception. Consequently, epistemic vigilance is a critical component of the argumentative theory of reasoning: for, if receivers were not vigilant, senders need not have the ability to display the coherence of their arguments in order to be able to convince receivers. Thus, any serious issues with epistemic vigilance as a concept would constitute a significant blow to the evolutionary arms race and consequently to the ATR as a whole. We return to the position of epistemic vigilance within the ATR in Section 4.2.3.

Let us now consider some points of disagreement between Kourken Michaelian (2013) and Sperber (and colleagues) and critically evaluate these problems to come to a conclusion about the plausibility and acceptability of Mercier and Sperber's theory.

4.2.2 What is epistemic vigilance exactly?

{sec:EV-def}

Sperber (2013) chalks up a considerable share of the disagreement between him and Michaelian to an alleged misunderstanding between the two authors on the exact definition of epistemic vigilance:

> Michaelian seems to attribute to us the view that 'epistemic vigilance is a matter of processes devoted to screening out incoming false information on the basis of available behavioural cues'. Showing that vigilance in this narrow sense is not efficient would, he holds, be quite damaging to our conjecture. This is a misunderstanding. (Sperber, 2013, p. 65)

While I agree with Sperber that Michaelian seems to attack a narrower version of epistemic vigilance, I do not blame Michaelian for the misunderstanding. Sperber and colleagues are (intentionally or unintentionally) vague in their original paper about exactly what epistemic vigilance is. Their flexible use of the notion of epistemic vigilance might not be inherently problematic, but given the central role epistemic vigilance plays in much of Sperber and Mercier's work, I believe we are long overdue an exact definition of epistemic vigilance. In this section, I will gather the details that together may constitute a definition of epistemic vigilance according to Sperber et al. (2010) and Sperber (2013).

Let us first consider the ontological status of epistemic vigilance. Although on the face of it, one may want to construe epistemic vigilance as a set of mechanisms for filtering incoming information, Sperber et al. (2010) describe humans as having evolved a 'suite of mechanisms' *for* – not *of* – epistemic vigilance. This leaves open the question of what epistemic vigilance itself could be. One candidate is a cognitive capability or skill, which seems to be supported by Mercier and Sperber (2011, p. 60) who describe epistemic vigilance as something that can be 'exercise[d to] some degree'. Moreover, Sperber et al. (2010, §5) import empirical evidence on the development of epistemic vigilance in children, which would point to vigilance being a capacity or skill as well.

A possibly useful perspective on the definition of epistemic vigilance comes from sleep science, a field of research related to neurology and neurophysiology. Schie et al. (2021) define vigilance *per se* as follows:

> Vigilance is defined as the capability to be sensitive to potential changes in one's environment, ie the capability to reach a level of alertness above a threshold for a certain period of time rather than the state of alertness itself. (p. 175)

It may not be immediately obvious how a definition from sleep science may at all be applicable in our attempt to define epistemic vigilance. However, I do believe that we may assume some degree of overlap between neurology and psychology, and that epistemic vigilance must relate in some way to vigilance *per se*.

All things considered, I believe epistemic vigilance would be best defined as *the capability to be sensitive to the trustworthiness of informants and communicated information*.

Next, let us consider the specifics of the processes in the 'suite of mechanisms for epistemic vigilance'. Just like Sperber and colleagues are somewhat vague about the ontological status of epistemic vigilance, they are rarely straight-forward about the exact mechanisms they consider to fall under the umbrella of epistemic vigilance. This, however, may not be as easily forgiven as their opacity surrounding epistemic vigilance's ontological status. Let us first consider the following quote from Sperber (2013):

> The probability of a biological trait evolving is contingent on its costs-benefits balance. Only if the benefits are greater than the costs, is it likely to evolve at all, and it is likely to evolve in a manner that, within the local range of possibilities, optimizes this balance. (Sperber, 2013, p. 62)

Following this emphasis on the importance of costs and benefits to the evolutionary account, one would expect Sperber (and colleagues) to provide a detailed analysis of the costs and benefits of epistemic vigilance. But since they do not explicate the contents of the suite of mechanisms for epistemic vigilance, we also do not get a clear picture of the cognitive costs of these mechanisms. Their costs-benefits analysis does not extend much beyond an argument of the form "epistemic vigilance has evolved in humans, therefore the costs of its mechanisms must not have outweighed its benefits", which begs the question. Michaelian (2013) addresses this issue by invoking dual systems to explicate the costs of epistemic vigilance. He proposes that epistemic vigilance contributes to the stability of communication only through type 2 processing. Interestingly, Sperber (2013) does not mention or address Michaelian's use of dual systems.

Is this begging the question or circular reasoning?

In short, the characterization of epistemic vigilance falls short of providing a convincing analysis of the costs involved in it, which is slightly puzzling considering Sperber's emphasis on the costs-benefits balance. We will see in Section 4.2.3 that the same issue arises when it comes to the benefits of epistemic vigilance.

Lastly, let us consider the relation between epistemic vigilance and reasoning. Sperber et al. (2010) describe reasoning as "a tool for epistemic vigilance, and for communication with vigilant addressees" (p. 378). This would imply that reasoning is an item in the suite of mechanisms for epistemic vigilance. However, in a similar way to how Sperber and colleagues are unspecific about epistemic vigilance's contribution to the stability of communication, they are unspecific about reasoning's contribution to epistemic vigilance.

All in all, Sperber and colleagues' definition and description of epistemic vigilance leaves a lot to be desired when it comes to the details. We return to this issue in Section 4.3 when we consider the broader picture of metatheoretical issues with the ATR as a whole.

4.2.3 Strong vs. weak readings of the argument for epistemic vigilance

{sec:strong-weak}

In their 2010 paper, Sperber and colleagues hint at different ways to interpret their argument — at different roles to attribute to epistemic vigilance:

> It is because of the risk of deception that epistemic vigilance may be not merely advantageous but indispensable if communication itself is to remain advantageous. (p. 360)

Sperber et al. seem to prefer to remain agnostic (or – put differently – vague) about the exact role of epistemic vigilance in explaining the stability of communication: is epistemic vigilance "merely advantageous", or is it "indispensable"? In other words, is it just the case that the benefits of epistemic vigilance outweigh its costs, leading to its having evolved in humans; or, stronger, would communication collapse if it were not for receivers' epistemic vigilance? These different readings of epistemic vigilance are also implicit in the following statement Sperber and colleagues make:

Add example

> People stand to gain immensely from communication with others, but this leaves them open to the risk of being accidentally or intentionally misinformed, which may reduce, cancel, or even reverse these gains. (p. 360)

If being misinformed reduces or cancels the benefits an addressee receives from the communicative event, then it would stand to reason that it would be advantageous for the addressee to be epistemically vigilant so as to maintain the positive benefits of communication. If misinformation however reverses the gains one receives from communication, then epistemic vigilance would be *necessary* for the receiver in order to not be negatively affected.

Kourken Michaelian picks up on these different views of epistemic vigilance in his 2013 response paper. He distinguishes between a strong and weak reading of Sperber et al.'s argument for epistemic vigilance, in line with this distinction between vigilance being indispensable or advantageous, respectively. He then outlines the assumptions that underpin each reading of the argument, and purports to show that these assumptions are unfounded, or at the very least too strong. According to Michaelian, the strong reading carries with it the assumption that dishonesty is sufficiently prevalent to necessitate vigilance on the side of the receiver; since, if epistemic vigilance is indispensable, non-vigilance must then yield a "dramatic reduction in the fitness of receivers" (Michaelian, 2013, p. 39). He presents empirical findings that maintain that lying is infrequent (Serota et al., 2010), and that therefore the strong reading of Sperber et al.'s argument cannot hold.

In his response to Michaelian, Sperber (2013) states that "I now believe that we could and should have been even less definite" (p. 63) in recognizing a stronger and weaker reading of their argument. Sperber argues that in the recognition of the two readings of the argument, one mistakenly regards communication as a static enterprise. The degree to which vigilance is advantageous or even indispensable to a receiver, varies greatly from situation to situation, he argues:

> The benefits of vigilance may be negligible in some communicative interactions and essential in other interactions. All I feel confident to say is that, without vigilance, human communication would be a very different and probably much more restricted affair. (Sperber, 2013, p. 63)

This conclusion, however plausible, leaves a lot to be desired when it comes to the details for the evolutionary story, and in particular for the picture of the evolutionary arms race. Recall Sperber's emphasis on the costs-benefits balance we discussed in Section 4.2.2. How can this importance ascribed to the costs-benefits balance be reconciled with his non-committal stance on how beneficial epistemic vigilance actually is?

4.2.4 Honesty or dishonesty as prior

{sec:honesty-dishonesty}

Ultimately, a fundamentally different outlook on human communication and cooperation seems to transpire from the accounts of Sperber, Mercier and colleagues on the one hand, and Michaelian on the other.

Sperber's 'evolutionary arms race' account takes dishonesty as prior. In short, dishonesty creates vigilance, vigilance then creates honesty, and honesty creates trust. As Sperber and colleagues write,

> We could not be mutually trustful *unless* we were mutually vigilant. (Sperber et al., 2010, p. 364)

In other words, vigilance is prior to trust. Vigilance is necessitated by dishonesty, and honesty and trust are contingent on vigilance. For this account to work, one must convincably argue for the evolutionary benefits of dishonesty; if dishonesty is the first step in the evolutionary arms race, it must be beneficial *per se*.

Michaelian, on the other hand, refutes this account and instead proposes that communication is stable just because speakers are honest; in other words, honesty is prior. This view is consistent with Michael Tomasello's account of the evolution of human cooperation, as we have seen in Section 2.4.1. In discussing how children altruistically share information with others, Tomasello writes

> Of course children soon learn to lie also, but that comes only some years later and presupposes preexisting cooperation and trust. If people did not have a tendency to trust one another's helpfulness, lying could never get off the ground. (Tomasello, 2009, p. 21)

In other words, trust is prior to deception: deception could not have emerged without trust. Analogously to Sperber's account, it then remains for Michaelian, Tomasello and the like to show how honesty is by itself beneficial.

As is often the case with 'chicken-or-egg' problems such as these, it could very well be that the truth lies somewhere in the middle. Perhaps the answer to the question is not as straight-forward as choosing between the options of 'we are fundamentally vigilant' and 'we are fundamentally trustful'. For now, I must say that the latter is more plausible than the former, due to Tomasello's

evolutionary account being a more coherent, complete and thus convincing one than Mercier and Sperber's arms race account.

4.2.5 Conclusion

In general, epistemic vigilance is convincing as a concept, wrapped in a compelling story. However, the details of this story are severely lacking, which raises some eyebrows. This is not to say that these details could not be added; but their theory is not as strong as the authors hold it to be. Continuing this line of thought, let us now zoom out to discuss the argumentative theory of reasoning metatheoretically.

4.3 Metatheoretical problems of the ATR

{sec:ont-atr}

One issue that we have seen come up in Section 3.2 as well as Section 4.2.2, is that Mercier and Sperber are oftentimes intentionally or unintentionally vague about particularities of their theory. For one, Mercier and Sperber (2009) leave a lot to be desired with regards to the ontological details of intuitive and reflective inference. There are moreover small differences between the way they characterize these concepts in different writings (Mercier and Sperber, 2009, 2011), which contributes to an overall impression of the authors not wanting to commit to specific definitions of their concepts.

Moreover, as we discussed in Section 4.2.2, Sperber et al. (2010) are imprecise on the ontological status of epistemic vigilance, the contents of the suite of mechanisms that contribute to vigilance, and the relation between reasoning and epistemic vigilance.

All things considered, the way in which Mercier and Sperber define and use their terminology is vague at its best and confusing at its worst. The problem may partly stem from the authors revising certain details of the ATR as it developed into the full-fledged theory as expounded in Mercier and Sperber (2017). For example, in their 2009 paper they explicate their dichotomy between intuitive and reflective inference, but slightly confusingly, they never use the term 'reflective inference' in Mercier and Sperber (2011) anymore. Moreover, in Mercier and Sperber (2009) the authors talk extensively about the mind's argumentation module, but in Mercier and Sperber (2017) there is no talk of such an argumentation module but a *reason* module instead. The reader is left to guess whether these two terms refer to the same concept or not.

This gives the appearance that something changed about Mercier and Sperber's views through the years. While this might very well not be the case, it does leave the audience to wonder.

All of this is of course understandable, as slight modifications and updates to one's theory constitute a natural part of the scientific process. However, Mercier and Sperber do not explicate these changes in terminology or in other details of their theory, which leaves the reader confused.

In general, the ATR is a convincing theory at first glance, but it fails to pass beyond this 'intuitively plausible' judgement due to the authors' slippery use of terminology, essential details that are missing from their argument, and the questionable assumptions that underly some of their claims.

Regarding this overall vagueness when it comes to details of the ATR, the following quote from Karl Popper's *Conjectures and Refutations* comes to mind:

> by making their interpretations and prophecies sufficiently vague they were able to explain away anything that might have been a refutation of the theory had the theory and the prophecies been more precise. In order to escape falsification they destroyed the testability of their theory. It is a typical soothsayer's trick to predict things so vaguely that the predictions can hardly fail: that they become irrefutable. (Popper, 1962, p. 37)

Popper's description of astrologers being able to explain away any counterarguments to their theory, is reminiscent of how it feels to read Mercier and Sperber's work. Though this phenomenological anecdote can hardly count as proof that their theory is unfalsifiable, it is remarkable. I believe that accusing Mercier and Sperber's theory of being unfalsifiable, and consequently unscientific, would be a step too far. Besides, I consider the metatheoretical analysis required for such a serious accusation to be out of scope for this thesis. All I feel confident to say is that the ATR has considerable metatheoretical issues that demand attention before we can take it seriously.

4.4 Conclusion

In summary, the argumentative theory of reasoning is an attractive and intuitively convincing theory, but ultimately fails to be much more than just that. Below the surface, the ATR lacks the detail needed to justify its debatable implicit assumptions. Moreover, the theory is quite disconnected from the broader context of human cooperation, which decreases its evolutionary plausibility. Lastly, the ATR hosts a number of metatheoretical issues, from which one might even be led to conclude that the theory is irrefutable in Popper's sense.

To end on an optimistic note, I do not believe that Mercier and Sperber's theory is unsalvageable. I do not see an obvious reason why the ATR's flaws could not be resolved by adding the necessary details and tethering it to a broader context. For now, though, the argumentative theory of reasoning is just not cutting it.

Conclusion

{ch:conclusion}

The aim of this thesis has been to critically evaluate Mercier and Sperber's argumentative theory of reasoning, which posits that the biological function of reasoning is to produce arguments and evaluate those of others.

To support this critical evaluation, we first had a look at evolutionary theory to provide some background to the evolutionary perspective of the ATR. We discussed biological and cultural evolution, causation in evolution, evolutionary psychology, terminological issues, and finally outlined a methodology for an evolutionary analysis of human communication. This evolutionary analysis of human communication saw us discuss definitions of communication, consider empirical work on communication in non-human animals and in human development, and (arguably most importantly) analyze the function of communication. We introduced the problem of the stability of communication, examined the evolution of human cooperation and how this relates to communication, and discussed lying and deception. Next, we laid out Mercier and Sperber's views in detail. We discussed their 'evolutionary arms race' of communication and their concept of epistemic vigilance, and saw how these components came together in the argumentative theory of reasoning. Finally, these findings culminated in a critical evaluation of the ATR, resulting in a couple of conclusions. Firstly, the function of human communication as it transpires from my research is not easily reconciled with the ATR as-is. The ATR's focus and context is too narrow, which makes it weak from an evolutionary perspective. Secondly, combining my own findings with criticisms by Michaelian (2013), I conclude that particular aspects of the epistemic vigilance story are unconvincing beyond a superficial glance. Thirdly, as a theory, the ATR is unsatisfactory because it is oftentimes vague or underspecified, and it lacks important details.

All in all, the argumentative theory of reasoning is intuitively attractive, but rests on implicit assumptions that are currently unconvincing. Moreover, the theory is plagued by vagueness and lacks important details. I hypothesize that the ATR is not unsalvageable however; but, there is definitely some substantial work needed before the ATR can really be taken seriously.

Further research

The ATR's issues naturally give rise to avenues for further investigation.

For starters, it would be nice to see Mercier and Sperber (or perhaps others) resolve the vagueness in the ontological aspects of the ATR, as we discussed in Section 4.3. This would require additional philosophical work, critically analyzing the ATR from a metatheoretical perspective.

Another interesting line of inquiry concerns the reconciliation of the ATR within a broader context. In Section 4.1, I argued that the ATR would be stronger if it were integrated within a broader evolutionary context of human cooperation à la Tomasello (2009), and conjectured that it would be possible to do so. The obvious next step is then to determine whether it is indeed the case that these two stories can be reconciled. This endeavor would be one of evolutionary anthropology mostly, with a philosophical touch. The open questions posed by the questionable assumptions discussed in Section 4.1 could perhaps be answered by findings from evolutionary anthropology. If this research would conclude that the ATR in its current version cannot be successfully integrated into a Tomasellan view on cooperation, then it could be fruitful to determine if the ATR could be modified in a such a way that it *can* be integrated.

The last avenue for further research I will mention here concerns a significant question that transpires from the discussion in Section 4.3: is the ATR unfalsifiable? To answer this question, one would carry out a detailed metatheoretical analysis into multiple aspects of the theory, such as the ontological and definitional rigor with which concepts are used, and how empirical work is imported to support the theory. Mercier and Sperber ascribe empirical believability to the ATR by formulating hypotheses that they argue to be entailed by the theory. It would be good to critically evaluate the status of these hypotheses against the whole theory, to assess whether they are indeed generated by the theory. Moreover, one could survey the field for evidence that refutes these hypotheses and evaluate its merit, to determine whether it could be the case that Mercier and Sperber are cherry-picking the empirical work they reference.

Concluding words

Now, let me end this thesis by ruminating on some broader implications.

I have personally experienced that viewing the particularities of the human experience through an evolutionary lens is an infinitely fascinating and rewarding endeavor. It allows us to see things in a broader context, especially since with each passing day our lived experiences move further from those of our ancestors. Focusing on our 'roots' in this way might even constitute some form of escapism, but it also allows us to understand ourselves and our place in the world better.

When I first read about the argumentative theory of reasoning, Mercier and Sperber's ideas felt optimistic to me. The ATR convinced me that we are not bad at reasoning; the so-called flaws of human reasoning are a feature, not a bug, so we are actually perfectly fine reasoners! Now that I have analyzed the theory in more detail though, I feel differently about the story they tell. Their focus on deception and vigilance has a more pessimistic flair (see also Section 4.2.4),

which is (seemingly) at odds with the general outlook on human evolution that Michael Tomasello emanates (cf. Tomasello (2009)). In the end, it is of course not the purpose of science to make us feel good about ourselves; but it is a pleasant side effect.

Bibliography

Allen, C. and M. Bekoff (1995). "Biological function, adaptation, and natural design". In: *Philosophy of Science* 62.4, pp. 609–622.

Andrews, K. (2015). *The animal mind: An introduction to the philosophy of animal cognition*. Routledge.

Apicella, C. L. and J. B. Silk (2019). "The evolution of human cooperation". In: *Current Biology* 29.11, R447–R450.

Ariew, A., R. Cummins, and M. Perlman (2002). *Functions: New essays in the philosophy of psychology and biology*. Oxford University Press, USA.

Ayala, F. J. (1999). "Adaptation and novelty: teleological explanations in evolutionary biology". In: *History and philosophy of the life sciences*, pp. 3–33.

Bateson, P. and K. N. Laland (2013). "Tinbergen's four questions: an appreciation and an update". In: *Trends in Ecology & Evolution* 28.12, pp. 712–718.

Benítez-Burraco, A., F. Ferretti, and L. Progovac (2021). "Human self-domestication and the evolution of pragmatics". In: *Cognitive Science* 45.6, e12987.

Benton, M. J., D. Dhouailly, B. Jiang, and M. McNamara (2019). "The Early Origin of Feathers". In: *Trends in Ecology & Evolution* 34.9, pp. 856–869. DOI: 10.1016/j.tree.2019.04.018.

Brinol, P. and R. E. Petty (2009). "Source factors in persuasion: A self-validation approach". In: *European review of social psychology* 20.1, pp. 49–96.

Bullinger, A. F., F. Zimmermann, J. Kaminski, and M. Tomasello (2011). "Different social motives in the gestural communication of chimpanzees and human children". In: *Developmental Science* 14.1, pp. 58–68.

Call, J. and M. Tomasello (2007). *The gestural communication of apes and monkeys*. Psychology press.

Carpenter, M., K. Nagell, M. Tomasello, G. Butterworth, and C. Moore (1998). "Social cognition, joint attention, and communicative competence from 9 to 15 months of age". In: *Monographs of the society for research in child development*, pp. i–174.

Chater, N. and M. Oaksford (2018). "The enigma is not entirely dispelled: A review of Mercier and Sperber's The Enigma of Reason". In: *Mind & Language* 33.5, pp. 525–532.

Cheney, D. L. and R. M. Seyfarth (1997). "Why animals don't have language". In: *Tanner lectures on human values*. Ed. by G. B. Peterson. Vol. 19. University of Utah Press, pp. 175–209.

Claidière, N., T. C. Scott-Phillips, and D. Sperber (2014). "How Darwinian is cultural evolution?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1642, p. 20130368. DOI: 10.1098/rstb.2013.0368.
Cosmides, L. (1989). "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task". In: *Cognition* 31.3, pp. 187–276.
Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. Grosset/Putnam, New York.
Dawkins, R. and J. R. Krebs (1978). "Animal Signals: Information or Manipulation?" In: *Behavioural Ecology: An Evolutionary Approach*. Ed. by J. R. Krebs and N. B. Davies, pp. 282–309.
Donahoe, J. W. (2003). "Selectionism". In: *Behavior theory and philosophy*. Ed. by K. A. Lattal and P. N. Chase. Springer, pp. 103–128. DOI: 10.1007/978-1-4757-4590-0_6.
Dor, D. (2017). "The role of the lie in the evolution of human language". In: *Language Sciences* 63, pp. 44–59.
Dutilh Novaes, C. (2018). "The enduring enigma of reason". In: *Mind & Language* 33.5, pp. 513–524.
Evans, J. S. B. (2003). "In two minds: dual-process accounts of reasoning". In: *Trends in cognitive sciences* 7.10, pp. 454–459.
Evans, J. S. B. and K. E. Stanovich (2013). "Dual-process theories of higher cognition: Advancing the debate". In: *Perspectives on psychological science* 8.3, pp. 223–241.
Evans, J. S. B. and D. E. Over (1996). *Rationality and reasoning*. Psychology Press.
Federle, M. J. and B. L. Bassler (2003). "Interspecies communication in bacteria". In: *The Journal of clinical investigation* 112.9, pp. 1291–1299.
Freeberg, T. M., K. E. Gentry, K. E. Sieving, and J. R. Lucas (2019). "On understanding the nature and evolution of social cognition: a need for the study of communication". In: *Animal Behaviour* 155, pp. 279–286.
Gilbert, D. T. (1991). "How mental systems believe". In: *American psychologist* 46.2, p. 107.
Goldstone, R. L., E. J. Andrade-Lotero, R. D. Hawkins, and M. E. Roberts (2024). "The emergence of specialized roles within groups". In: *Topics in Cognitive Science* 16.2, pp. 257–281.
Gómez, J. C. (2004). *Apes, monkeys, children, and the growth of mind*. Harvard University Press.
Grice, H. P. (1975). "Logic and conversation". In: *Speech acts*. Brill, pp. 41–58.
Hare, B. (2017). "Survival of the friendliest: Homo sapiens evolved via selection for prosociality". In: *Annual review of psychology* 68.1, pp. 155–186.
Hertwig, R. and G. Gigerenzer (1999). "The 'conjunction fallacy'revisited: How intelligent inferences look like reasoning errors". In: *Journal of behavioral decision making* 12.4, pp. 275–305.
Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press. DOI: 10.4159/9780674985155.
Hrdy, S. B. (2009). *Mothers and others: The evolutionary origins of mutual understanding*. Harvard University Press.

Hume, D. (1739/1978). *A treatise of human nature*. Oxford University Press.
Iverson, J. M. and S. Goldin-Meadow (2005). "Gesture paves the way for language development". In: *Psychological science* 16.5, pp. 367–371.
Johnson, M. R. (2005). "Historical Background to the Interpretation of Aristotle's Teleology". In: *Aristotle on Teleology*. Oxford University Press, pp. 15–39. DOI: 10.1093/0199285306.003.0002.
Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
Khait, I., O. Lewin-Epstein, R. Sharon, K. Saban, R. Goldstein, Y. Anikster, Y. Zeron, C. Agassy, S. Nizan, G. Sharabi, et al. (2023). "Sounds emitted by plants under stress are airborne and informative". In: *Cell* 186.7, pp. 1328–1336.
Koreň, L. (2023). "Have Mercier and Sperber untied the knot of human reasoning?" In: *Inquiry* 66.5, pp. 849–862.
Lachmann, M., S. Szamado, and C. T. Bergstrom (2001). "Cost and conflict in animal signals and human language". In: *Proceedings of the National Academy of Sciences* 98.23, pp. 13189–13194.
Laland, K. N. and G. R. Brown (2002). *Sense and nonsense: Evolutionary perspectives on human behaviour*. Oxford University Press, USA.
Laland, K. N., J. Odling-Smee, W. Hoppitt, and T. Uller (2013). "More on how and why: cause and effect in biology revisited". In: *Biology & Philosophy* 28, pp. 719–745.
Lee, K. (2013). "Little liars: Development of verbal deception in children". In: *Child development perspectives* 7.2, pp. 91–96.
Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.
Lipton, P. (2009). "Causation and Explanation". In: *The Oxford Handbook of Causation*. Ed. by H. Beebee, C. Hitchcock, and P. Menzies. Oxford University Press. DOI: 10.1093/oxfordhb/9780199279739.003.0030.
Maynard Smith, J. (May 1994). "Must reliable signals always be costly?" In: *Animal Behaviour* 47.5, pp. 1115–1120. ISSN: 0003-3472. DOI: 10.1006/anbe.1994.1149. URL: http://dx.doi.org/10.1006/anbe.1994.1149.
Mayr, E. (1961). "Cause and effect in biology". In: *Science* 134.3489, pp. 1501–1506.
McNamara, J. M. and F. J. Weissing (2010). "Evolutionary game theory". In: *Social behaviour: genes, ecology and evolution*, pp. 88–106.
Meibauer, J. (2018). "The linguistics of lying". In: *Annual Review of Linguistics* 4.1, pp. 357–375.
Mercier, H. and D. Sperber (2009). "Intuitive and reflective inferences". In: *In two minds: Dual processes and beyond*. Ed. by J. Evans and K. Frankish.
— (2011). "Why do humans reason? Arguments for an argumentative theory". In: *Behavioral and Brain Sciences* 34.2, pp. 57–74.
— (2017). *The enigma of reason*. Harvard University Press. DOI: 10.4159/9780674977860.
Michaelian, K. (2013). "The evolution of testimony: Receiver vigilance, speaker honesty and the reliability of communication". In: *Episteme* 10.1, pp. 37–59.
Millstein, R. L. (2021). "Genetic Drift". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University.

Moshman, D. and M. Geil (1998). "Collaborative reasoning: Evidence for collective rationality". In: *Thinking & Reasoning* 4.3, pp. 231–248.
Novaes, C. D. (2020). *The dialogical roots of deduction: Historical, cognitive, and philosophical perspectives on reasoning*. Cambridge University Press.
Oaksford, M. and N. Chater (1994). "A rational analysis of the selection task as optimal data selection." In: *Psychological review* 101.4, p. 608.
Penn, D. J. and S. Számadó (2020). "The Handicap Principle: how an erroneous hypothesis became a scientific principle". In: *Biological Reviews* 95.1, pp. 267–290.
Popper, K. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.
Quine, W. V. O. (1960). *Word and object*.
Rohwer, S. and F. C. Rohwer (1978). "Status signalling in Harris sparrows: experimental deceptions achieved". In: *Animal behaviour* 26, pp. 1012–1022.
Saul, J. M. (2012). *Lying, misleading, and what is said: An exploration in philosophy of language and in ethics*. Oxford University Press.
Schie, M. K. van, G. J. Lammers, R. Fronczek, H. A. Middelkoop, and J. G. van Dijk (2021). "Vigilance: discussion of related concepts and proposal for a definition". In: *Sleep Medicine* 83, pp. 175–181.
Schliesser, E. (2019). "Synthetic philosophy". In: *Biology & Philosophy* 34.2, pp. 1–9.
Schouls, P. A. (1972). "Descartes and the Autonomy of Reason". In: *Journal of the History of Philosophy* 10.3, pp. 307–322.
Scott-Phillips, T. C. (2008). "On the correct application of animal signalling theory to human communication". In: *The evolution of language*. World Scientific, pp. 275–282.
— (2015). "Nonhuman primate communication, pragmatics, and the origins of language". In: *Current Anthropology* 56.1, pp. 56–80.
— (2018). "Cognition and communication". In: *The International Encyclopedia of Anthropology*. Ed. by H. Callan and S. Coleman. John Wiley & Sons.
Scott-Phillips, T. C., K. N. Laland, D. M. Shuker, T. E. Dickins, and S. A. West (2013). "The niche construction perspective: a critical appraisal". In: *Evolution* 68.5, pp. 1231–1243.
Serota, K. B., T. R. Levine, and F. J. Boster (2010). "The prevalence of lying in America: Three studies of self-reported lies". In: *Human Communication Research* 36.1, pp. 2–25.
Seyfarth, R. M. and D. L. Cheney (2003). "Signalers and receivers in animal communication". In: *Annual review of psychology* 54.1, pp. 145–173.
Seyfarth, R. M., D. L. Cheney, and P. Marler (1980). "Vervet monkey alarm calls: semantic communication in a free-ranging primate". In: *Animal Behaviour* 28.4, pp. 1070–1094.
Sherman, P. W. (1977). "Nepotism and the Evolution of Alarm Calls: Alarm calls of Belding's ground squirrels warn relatives, and thus are expressions of nepotism." In: *Science* 197.4310, pp. 1246–1253.
Sloman, S. A. (1996). "The empirical case for two systems of reasoning". In: *Psychological bulletin* 119.1, p. 3.

Sperber, D. (2001). "An Evolutionary Perspective on Testimony and Argumentation". In: *Philosophical Topics* 29.1/2, pp. 401–413.
— (2013). "Speakers are honest because hearers are vigilant: Reply to Kourken Michaelian". In: *Episteme* 10.1, pp. 61–71.
Sperber, D., F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson (2010). "Epistemic vigilance". In: *Mind & language* 25.4, pp. 359–393.
Sperber, D. and D. Wilson (1986). *Relevance: Communication and cognition*. Vol. 142. Harvard University Press Cambridge, MA.
Stanley, S. (2021). "Cultural evolutionary theory and the significance of the biology-culture analogy". In: *Philosophy of the Social Sciences* 51.2, pp. 193–214.
Sterelny, K. (2018). "Why reason? Hugo Mercier's and Dan Sperber's The Enigma of Reason: A New Theory of Human Understanding". In: *Mind & Language* 33.5, pp. 502–512.
Tinbergen, N. (1963). "On aims and methods of ethology". In: *Zeitschrift für Tierpsychologie* 20.4, pp. 410–433.
Tomasello, M. (2005). *The Emergence of Social Cognition in Three Young Chimpanzees*.
Tomasello, M. (2008). *Origins of human communication*. MIT Press. DOI: 10.7551/mitpress/7551.001.0001.
— (2009). *Why we cooperate*. MIT Press.
Tversky, A. and D. Kahneman (1983). "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." In: *Psychological review* 90.4, p. 293.
Uller, T. and K. N. Laland (2019). *Evolutionary causation: biological and philosophical reflections*. Vol. 23. MIT Press.
Vorobeychik, Y., Z. Joveski, and S. Yu (2017). "Does communication help people coordinate?" In: *PloS one* 12.2, e0170780.
Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. John Wiley & Sons.
Walter Jr, O. M. (1951). "Descartes on reasoning". In: *Speech Monographs* 18.1, pp. 47–53.
Wason, P. C. (1968). "Reasoning about a rule". In: *Quarterly journal of experimental psychology* 20.3, pp. 273–281.
Wells, J. C. (2006). "The evolution of human fatness and susceptibility to obesity: an ethological approach". In: *Biological Reviews* 81.2, pp. 183–205.
Williams, B. (2002). *Truth and truthfulness: An essay in genealogy*. Princeton University Press.
Zahavi, A. (1975). "Mate selection — a selection for a handicap". In: *Journal of theoretical Biology* 53.1, pp. 205–214.
Zahavi, A. and A. Zahavi (1999). *The handicap principle: A missing piece of Darwin's puzzle*. Oxford University Press.
Zuberbühler, K., R. Noë, and R. M. Seyfarth (1997). "Diana monkey long-distance calls: messages for conspecifics and predators". In: *Animal Behaviour* 53.3, pp. 589–604.