Born to argue?

The argumentative theory of reasoning revisited

Flip Lijnzaad
Supervisor: Karolina Krzyżanowska

September 23, 2024

Contents

Introduction

{ch:introduction}

Communication is one of the most fundamental parts of the human experience: it's hard, nigh impossible, to imagine life 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 our ability to communicate more than just the truth-conditional content of our sentences. More often than not, the communicated content of the sentences we utter extends way beyond the words we speak. For example, take the following exchange (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, pointing at 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?

As a small aside: much of the pragmatics literature uses the terms reasoning and inference interchangeably (see for example Stephen Levinson's usage of the term 'reasoning' in Levinson (1983, p. 218)). However, Mercier and Sperber do sharply distinguish between reasoning and inference (see the very first paragraph of Mercier and Sperber (2011)). 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.

In 2011, Hugo Mercier and Dan Sperber proposed a revolutionary theory of reasoning, in response to a puzzling problem on the function of reasoning. It has been a longstanding view in philosophy that the function of reasoning is to enhance an individual's knowledge. However, experimental findings from the psychology of reasoning apparently show again and again that humans are not good at reasoning. This raises the question of how this can be the case, evo-

lutionarily speaking; why would reasoning have evolved to serve its function poorly?

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

Mercier and Sperber buttress the 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 perfectly 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) or their evolutionary framework underpinning the theory (Dutilh Novaes, 2018) to the cognitive-scientific framework underpinning their 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 communication.

In order to do so, 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, and in particular the function of communication. These findings will 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 it, 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 considers human communication from an evolutionary perspective, following the methodological steps outlined in the previous chapter. In particular, this chapter will see us discussing at length the function of communication, which will provide us with some important anchor points for scrutinizing the ATR.

Then in Chapter 3, I will expound the argumentative theory of reasoning by discussing a number of foundational papers for the theory. All of this 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 that became apparent from 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, Chapter 5 recounts the whole story and highlights avenues to continue this research.

1 | The chicken and the egg: the evolutionary approach

{ch:evolution}

Before we are able to answer any *why*-questions about the evolution of reasoning and communication, some groundwork needs to be laid out. For what are the processes underlying evolution, and how does evolutionary causation work? What does it mean for some trait to evolve 'for the purpose of' another trait? Are we even justified in using this kind of terminology when it comes to evolution? And what intermediate questions will we need to ask ourselves in order to ultimately answer the question of why we reason and communicate?

This chapter serves to answer these and related questions. It by no means provides 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 to touch on a number of issues in the field that are relevant to our endeavor, such that we have a stronger foundation for our investigation into the evolution of reasoning and communication.

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 consecutive 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 we will consider how the process of cultural evolution fits this definition of selection.

[1] See 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'; 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.

Firstly, in standard evolutionary theory the processes introducing variation are mutation (changes in an organism's DNA) and migration (the movement of (genetic material of) organisms from one population to another) (Scott-Phillips et al., 2013). Variation by itself is undirected; it is only due to sorting that the selection process becomes directed (Donahoe, 2003).

Too many brackets

Add a bit more on sorting: see Sperber et al. (2010) notes for an example. Try to map it to what people already know, because they're probably familiar with the process, but just not with the term.

Make sure you yourself are clear on the difference between the higher-level process of selection, and the lower-level process of sorting.

Secondly, in standard evolutionary theory the sorting process amounts to natural selection and genetic drift. Natural selection acts upon the variation introduced by mutation and migration such that the genes that enhance an organism's fitness persist over time in the population (Scott-Phillips et al., 2013). 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. Genetic drift, on the other hand, is a *random* sorting process, resulting from a sampling error due to populations being finite in size[3].

Do some research on genetic drift and write more about it. Biologists seem to disagree on how important genetic drift is compared to natural selection; mention that. And the definition as it is right now, is not understandable: unpack it a bit, mention its importance is debated, and keep it aside.

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.

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 underlying each of the three steps of the selection process.

Firstly, variation in culture is introduced by error and by innovation. Secondly, sorting of behaviors and habits in culture can happen through two different routes. A behavior can be sorted (selected for) 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. Also, a behavior or habit may be sorted in a more 'classic' evolutionary way: a habit may be selected for 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 . Thirdly, culture may be retained through cultural inheritance, that is, through mechanisms of social learning .

Add example

Add example

Add example

Add example

[3]See Millstein (2021) for discussion on the concept of genetic drift and how it compares with natural selection.

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 et al. (2014) for discussion and a formal account of cultural evolution. 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 reasoning and communication 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 the same. Thus, for the remainder of this thesis, I will use the term 'evolution' to talk about the development of characteristics (in our case, cognitive capacities) in humans over time, remaining agnostic about whether this development is due to biological evolution or cultural evolution.

Try to Find this reference

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; let us 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).

Evolutionary causation is a subfield of philosophy of biology that has continued to see widely diverging opinions (Baedke, 2021; Scott-Phillips et al., 2013). In this section I will restrict focus to four topics in evolutionary causation that are of interest to this thesis. First, I will discuss the distinction between proximate and ultimate causation. Then, I will briefly cover Tinbergen's (1963) questions for explaining animal behavior, which will be discussed at length in Section 1.5. Thirdly, we will have a look at niche construction and reciprocal causation. Lastly, we circle back to our research question and discuss what it entails for one trait to evolve for the purpose of another.

In evolutionary causation, one may distinguish *proximate* from *ultimate* causes. Proximate causes are the immediate influences on a trait: they explain how the trait results from the internal and external factors causing it. Ultimate causes, on the other hand, provide the higher-level historical and evolutionary explanation of those traits (Mayr, 1961). In other words, these two different causes relate to two different explanatory 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[4], one needs to answer both of these explanatory questions about a trait in order to obtain a complete understanding of it.

Add an example to illustrate

In a seminal proposal considered by some to be an extension of Mayr's di-

[4]See (Laland et al., 2013) for a discussion of the exact origins of the distinction.

chotomy (Laland et al., 2013), 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 proximate causation of the behavior, (2) the lifetime development of the behavior, (3) the function[5] of the behavior, and (4) the evolutionary history of the behavior. Since Tinbergen, 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 'function' and 'evolution' questions as ultimate questions (*why*-questions) (Bateson and Laland, 2013; Laland et al., 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 inform the methodology for this thesis and will thus discuss it in greater detail in Section 1.5.

According to standard evolutionary theory, evolution is a causally *unidirectional* affair: natural selection shapes an organism in such a way that the organism is better adapted to its environment, and as such, the causal chain starts with the environment and ends with the organism. In recent years however, biologists have been looking beyond this unidirectional view and are starting to consider the role of *reciprocal* causation in evolution. According to this concept, not only does the environment cause changes in an organism through evolutionary processes, the organism also causes changes in its environment through its actions:

> To varying degrees, organisms choose habitats and resources; construct aspects of their environments, such as nests, holes, burrows, webs, pupil cases, and a chemical milieu, and destroy other components; and frequently choose, protect, and provision nursery environments for their offspring (Day et al., 2003, p. 81)

This process by which organisms influence their environment is referred to as *niche construction*. A prominent example of niche construction is the evolution of lactose tolerance in adult humans. The prevalence of lactose tolerance in a population correlates with whether that population has a history of dairy farming. This suggests that due to their adoption of dairy farming, the individuals in the population have come to rely upon their tolerance of lactose, and this reliance amounts to a selection pressure for lactose tolerance into adulthood (Scott-Phillips et al., 2013). Much discussion remains on the exact role that niche construction should play in our evolutionary theories. Proponents of *niche construction theory* maintain that niche construction is a process that operates alongside natural selection and is an evolutionary factor in its own right, whereas skeptics argue that standard evolutionary theory can account for niche construction effects (Scott-Phillips et al., 2013). However, the concept of niche construction itself is uncontroversial: there is plenty of empirical evidence for the fact that organisms may partake in shaping their niche (Scott-Phillips et al., 2013).

[5] See Section 1.4 for a discussion of function in biology.

Lastly, we briefly touch on what it means for one trait to evolve because of, or for the purpose of, another trait. In this thesis, we are interested in what evolutionary benefits one feature (reasoning) may have to another feature (communication). In order to answer this question, we should first consider what the function[6] of communication is to humans; only then can we consider whether the function of reasoning could be to advance communication and in doing so, improve the fitness of humans.

Don't overuse parentheses

Reasoning vs. the ability to reason: delve into this somewhere. Behavior vs. the ability to exhibit behavior. The convention in the evolutionary literature is to talk about these behaviors at a higher level of abstraction: mention this somewhere.

This last paragraph deserves more discussion

Missing: concluding remarks about what I will do with the concepts from this section

1.3 Evolutionary psychology

{sec:evol-psych}

In order to answer the question of whether reasoning evolved for the purpose of communication, we will also need to zoom out to consider the field of evolutionary psychology as a whole. What is the merit, and the validity, of adopting an evolutionary approach in our endeavor at all?

Weird place to ask this question: should be moved to the introduction because it's so basic, and it motivates the research question. Here, mostly a matter of phrasing: this section is more specific than just asking "what is the merit of considering evolution at all". So rephrase this.

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; see Laland and Brown (2002, pp. 19–20) for an overview. 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 the 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 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).

[6] See Section 1.4.

Are the groups the EEA or the savannah? Elaborate more on this: is it the physical environment or does it include the groups? Aren't the groups a trait that evolved? Think about this

The assumptions underlying the use of the concept of an EEA are that (1) 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; and (2) 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 purpose. Secondly, we do not know enough about the process of evolution to make assumption (2); 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). Thirdly, 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. Lastly, the EEA argument does not take into account reciprocal causation or niche construction.

Probably good to mention the nature of this evidence, because this evolutionary timeline point is important

Elaborate on this last issue, because it seems to be important for my argument. Also, it seems like a strange point to make, because especially with humans, it seems obvious that niche construction is a thing that might play a role. Why would the EEA concept rely on assumptions of unidirectional causation?

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 is that humans throughout history have been dependent on cooperation and strong social groups for survival.

This paragraph deserves some attention: some more stuff about cooperation, and how social bonds and cooperation can be beneficial. Because sharing food as-is is not beneficial per se. The issue with the "sharing food is beneficial" could also be resolved by adding half a sentence of explanation about social bonds being beneficial, but I'd like to be more rigorous.

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 often when hunting does succeed the yield is large; sometimes even too large for the hunter and his relatives. In this case, food sharing within or between groups is beneficial. Another way that early humans counterbalanced the variable returns of hunting was to also rely on gathering plant foods, which yielded more predictable returns. In this case cooperation through shared labor was also beneficial, since some foods required com-

plex foraging techniques to acquire it, or required complex processing (through 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).

Add also some things from Freeberg et al. (2019)?

The following paragraph might warrant a larger discussion about domain-specificity; see commented out comment. Right now, it's not clear that this is relevant, so delve more into this if it's relevant (and move it to the relevant spot, probably). Else, remove it

Another topic of discussion in evolutionary psychology that is of importance to our investigation is that of domain-specificity of the psychological mechanisms. The argument has been made that these adaptive mechanisms are necessarily problem- or domain-specific, because the evolutionary process would not favor general solutions to specific problems (Buss, 2015, p. 50). However, as with the EEA, issues with this stance have been raised: the push to domain-specificity can be said to rely on overly strong assumptions about the modularity of the brain; and moreover, there is also a push to domain-generality of cognitive skills because domain-general skills are neurologically more cost-efficient than domain-specific skills (Laland and Brown, 2002).

1.4 Teleological notions in evolutionary theory

{sec:teleology}

This section needs quite some work: see notes from discussion with Karolina

Next, it is important to scrutinize the terminology that I will be using throughout this thesis. 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 *function*, *purpose*, or *utility* of a trait. At first glance, the usage of these terms in discussing evolution would seem to be inappropriate; for evolution is a process of nature, not purposefully performed by an agent, and it is thus without any intentionality or goals. And indeed, this teleological terminology has its roots in pre-Darwinian conceptions of nature: it originates from Aristotle's views on causation, and it was subsequently adopted by creationist Muslim and Christian scholars (Johnson, 2005).

Maybe reformulate this sentence again: still too convoluted?

Reformulate this reference to Aristotle: metaphysics/nature more than causation. Be safe (i.e. rather broad than specific) about the phrasing here to not upset historians of philosophy. "Aristotle's views on nature" is probably best

In general, teleological explanations in biology are quite controversial: not only is the usage of the specific terminology itself debated (Ayala, 1999, p. 27 and references therein), the concept has been criticized for its apparent lack of formalization and insufficient argumentative persuasiveness (Baedke, 2021, p. 83).

Address the controversy around teleological explanations. Talk about instrumentalism, usefulness of the concepts. Lack of formalization is not such a big problem for the purpose here maybe, but the other thing is more of a problem. Address why they won't be a problem for you. Can mention that MS assume it as well, this teleological explanation is at the heart of their thesis (quote it?), so it's their problem to defend this. I work using the same assumptions as them.

However, explanations in terms of goals and function have considerable instrumental value in describing evolutionary processes. Throughout this thesis, I will be adhering to the conception of teleological explanations of Ayala (1999), which is as follows:

> 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*. (p. 13)

In this respect, the evolutionary process of adaptation merits a teleological explanation: the function of a trait (its 'contribution to a specific property or state of the system') is the reason that the trait exists, because it exists as a consequence of natural selection.

Add example here

Is this view compatible with reciprocal causation and niche construction? I think so; they're a complication for the whole picture, not necessarily for using this definition.

Is this view compatible with cultural learning? From the quote, it doesn't necessarily follow that it's about biology necessarily. Think about this, and after writing a section on culture, state to what extent and in what way we'll adhere to Ayala (1999)

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. The proximate end is then the 'immediate' function the feature serves, and the ultimate end is the reproductive success of the organism.

Add example

A footnote to this account is that not all features of organisms can be explained teleologically; only if the feature has arisen and persisted as a direct result of natural selection, a teleological explanation is in place.

This "direct result of natural selection" is very vague/slippery; acknowledge this, and elaborate more on it if it turns out to be important for my thesis. A way to do this would be to contrast it with an indirect result. Talk about side effects?

1.5 Adopting and adapting Tinbergen's four questions

{sec:tinbergen}

In this section: terminology issue: trait vs. feature vs. characteristic vs. behavior. Address this earlier on in the chapter, it also relates in a way to the ability to exhibit behavior vs. the behavior itself.

As mentioned in Section 1.2, Tinbergen (1963) proposed an influential framework of problems[7] that should be addressed if one intends to give a complete

[7] In the literature (e.g. Bateson and Laland (2013), the terms 'problems' and 'questions' are used

account of a behavior an animal exhibits. Let us dive more deeply into Tinbergen's framework here, as it will turn out to form a desirable foundation for the current investigations.

As mentioned before, the four problems that Tinbergen argued to be central to the study of behavior are causation, survival value, ontogeny, and evolution. Although these problems were originally raised in regards to animal behavior, the framework has since been adopted for analyzing the characteristics of organisms in general, and can even be used to gain understanding of nonliving systems, such as traffic lights (Bateson and Laland, 2013).

This last sentence is too ambiguous in its focus: rephrase so that focus is more on characteristics and less on organisms

I will now discuss each of these problems in more detail, such that we can ultimately come to a set of methodological questions to guide us in investigating human communication and reasoning.

1.5.1 Causation

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 and reasoning 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 communicative and reasoning behavior. Although it has been emphasized that we can only gain a full understanding of a behavior if the four problems are addressed simultaneously (Bateson and Laland, 2013; Tinbergen, 1963), I believe I am justified in leaving the proximate-causation problem for future research.

1.5.2 Survival value

The second problem that Tinbergen outlines relates to the value a behavior provides to an animal's survival: how does the behavior contribute 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 characteristic's function can change over time: the *current* utility that a characteristic 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). We will discuss in Chapter 2 and **??** what can be construed as the original and current utilities or functions of the cognitive capacities we are dealing with.

Try to add also example in humans of original and current utility not lining up: fat retention?

Possibly change this comment after Chapter 2 and 3 are more or less finished

interchangeably. I will take 'problems to address' and 'questions to answer' to be synonymous, and will use these terms interchangeably.

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 that 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 characteristic, which denotes the way the characteristic contributes to the fitness of the organism. This leads us to the following formulation of Tinbergen's question for our purposes:

(1) a. What was the original utility of communication to humans? And what is the current utility of communication to humans?

b. What was the original utility of reasoning to humans? And what is the current utility of reasoning to humans?

The distinction is pretty relevant, but rigorous discussion of both utilities won't be necessary: only use the distinction, don't discuss it. This distinction might be an avenue of scrutiny for Mercier & Sperber

1.5.3 Ontogeny

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.

This section is very short, but I don't feel like anything can/needs to be added?

So this leads us to the following question:

(2) a. How does the capacity for communication develop throughout childhood?

b. How does the capacity for reasoning develop throughout childhood?

1.5.4 Evolution

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 close 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 the 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 these historical processes, so we will only take up the latter of these two questions.

This leads us to the following formulation of Tinbergen's evolutionary question:

(3) a. What is the evolutionary history of human communication? How can its evolutionary trajectory from our nearest evolutionary relatives to us be explained?

b. What is the evolutionary history of human reasoning? How can its evolutionary trajectory from our nearest evolutionary relatives to us be explained?

1.5.5 A fifth Tinbergen question: observation and description

Emphasize the importance of this question here briefly, yes: but the definition should already be in the introduction, since the concepts are mentioned in the RQ and the title of the thesis. So drop this header, and probably the questions as well

A problem that is mentioned by Tinbergen in his original paper 1963, but not included as one of the core problems of his framework, and more or less never included by authors discussing his framework (Allen and Bekoff, 1995; Laland and Brown, 2002; Laland et al., 2013), is the problem of *describing* the observed behavior. In the case of describing reasoning and communication, this issue is akin to the problem of defining and delineating what we take to be reasoning and what we take to be communication. This is by no means a trivial issue, which is what warrants its inclusion in the set of questions we will ask ourselves in thesis:

(4) a. What is human communication?

b. What is human reasoning?

1.6 Conclusion

{sec:evo-conclusion}

Throughout this chapter, it has become apparent that for none of the topics in evolutionary theory discussed here consensus has been reached 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 to account in detail for their shortcomings. Let us conclude this chapter by first gathering four key assumptions that will inform this thesis, and then restating the methodological questions that will guide its investigations.

The first one 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 order to answer the question of whether reasoning evolved for communication, we must consider not only reasoning but also communication in detail. This is because for reasoning to have evolved for the purpose of communication, the latter must have been evolutionarily advantageous in its own right, such that advancements in reasoning could have advanced communication to such an extent that it made communication more evolutionarily advantageous. Moreover, as we will see, a thorough investigation of communication will illuminate the role reasoning plays in communication.

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. This assumption is especially important in the analysis of human communication.

The fourth and last 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 both kinds of evolution have the same underlying mechanisms of selection.

Lastly, regarding the questions we will ask ourselves in the following two chapters: the discussion in Section 1.5 has yielded four questions reformulated and adapted from Tinbergen's (1963) framework. Here, I restate these questions in a general manner and in an order that will be most useful to the investigations in Chapter 2 and **??**.

Definition	How can this cognitive capacity be defined and delineated?
Development	How does this cognitive capacity develop throughout childhood?
Evolution	What is the evolutionary history of this cognitive capacity; how can its evolutionary trajectory from our nearest evolutionary relatives to us be explained?
Utility	What are the original and current utilities of this cognitive capacity to humans?

Now that we have gathered the assumptions and questions, it is time to consider the cognitive capacity that is primary in the context of our research question: communication.

2 | On communication

{ch:communication}

Human communication is uniquely sophisticated and remarkably pervasive. In this chapter, we will dive head-first into studying communication, 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. Within this discussion, we also consider the evolution of human cooperation, the issue of the stability of communication, and the role of deception in this.

This introduction is a bit stiff; change after Introduction & Ch. 1 are finished

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 an process necessarily involving a sender (communicator or speaker) and at least one receiver (addressee or listener). In the literature, especially that on animal communication, the terms 'communication' and 'signalling' are used interchangeably; therefore, I will consider them to be interchangeable as well.

Some authors regard communication to inherently be a tool of persuasion, which then translates to their very definition of communication: for example, on the manipulative model of communication, communication can be taken to occur

> 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)

Here we see Dawkins and Krebs embedding a teleological explanation in their definition, when they speak of "the result of selection"; recall Section 1.4. But more pressingly, they also include the aspect of benefits to the sender and thus some stance on the function of communication in it ("to the advantage of the [sender]"). Because of this, I believe this definition to be insufficiently parsimonious in its assumptions. So let us instead turn to some conceptions of com-

munication that are more neutral with regards to how communication may be used.

In their discussion of communication as it relates to social cognition, Freeberg et al. (2019) 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)

This is a very broad conception of communication, and it encompasses more cases than Dawkins and Krebs's definition. It not only covers our case at hand – humans speaking and listening to each other – but also 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), plants communicating distress through airborne sounds (Khait et al., 2023), and many more cases. One notable case in particular is that of the peacock's tail — a characteristic of the peacock that influences the behavior of other individuals, in the sense that it makes him more sexually attractive to peahens. We return to this example in Section 2.4.2.

Scott-Phillips (2015) contrasts two different models of communication with each other: the classical *code model* of communication, and the *ostensive-inferential* model of communication. 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.

However, in order to capture human communication, the code model is too simplistic, 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); consider also Quine (1960)'s views on radical translation). Therefore, a move away from the code model of communication towards the *ostensive-inferential* model of communication would be in order. This model, introduced by Sperber and Wilson (1986) as part of their neo-Gricean framework of relevance theory, takes into account the intentionality inherent in human communication.

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 their receiver with an *ostensive* behavior. If their receiver receives their communicative intention, then ostensive-inferential communication has occurred.

Add example because this is very vague now

[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.

Currently, there is no evidence that any species other than humans communicate ostensively (Scott-Phillips, 2018). As a result, not only may one distinguish between the code model and the ostensive-inferential model to define what communication entails, one may also conceptualize these two models 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, I would like to note 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 (Sperber and Wilson, 1986) and in their writings, Mercier and Sperber often refer to the principles of relevance theory (Mercier and Sperber, 2009, 2011; Sperber et al., 2010).

2.2 Communication in non-human animals

{sec:comm:phylogeny}

Now we turn to the evolutionary question: what is the evolutionary history of communication, and more specifically, how can the trajectory from primate communication to human communication be explained?

As already mentioned, one fundamental difference between the communication of non-human animals and humans is by which model their communication is best described: the code model and the ostensive-inferential model, respectively (Scott-Phillips, 2015, 2018).

Communication is used by non-human animals for a wide range of purposes, and it can be elicited by a number of 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.

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 using it to signal 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:

> [information acquired by listeners] may include, but is not limited to, 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)

The model that best describes them is a *consequence* of the difference between the two: so the difference is that which makes one model work better for NHA and the other better for humans. Reformulate

Connect this to your definition!

Give example / elaborate (see Tomasello (2008)

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 then refer to stimuli received through the external senses, such as visual, auditory and olfactory senses. For example, if I stop petting my dog (sensory stimulus), she will direct her gaze at me (communicative behavior) to indicate that she would like me to continue. Mental stimuli on the other hand can be viewed as the mental states an animal attributes to another animal. For example, . This 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, except for possibly chimpanzees (Seyfarth and Cheney, 2003).

Give different example, with reference

Address the controversy surrounding theory of mind; see Andrews (2015)

Add example from the literature

Tomasello (2009) notes that in the case of pointing gestures, humans use these mainly to be informative (i.e. cooperative), whereas primates use this gesture mainly or perhaps even exclusively for imperative motives. Experimentally, Tomasello and colleagues 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 (Bullinger et al., 2011). As Tomasello (2009) concludes, "[the infants] could not help but be informative" (p. 17). Speaking of which, let us now consider communication in human infants.

As we will discuss deception in communication a great deal (for example in Section 2.4.3), let us briefly touch on deception in animals as well. As noted by Dor (2017) in his discussion of deception, primates possess the ability to deceive others. However, compared to humans, their ability to deceive is rather primitive. Dor argues that this is due to the fact that language uniquely enables humans to communicate with each others' imaginations. 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 themself. However, their ability to fabricate information is very limited.

Example or elaborate

2.3 Communication in children's development

{sec:comm:ontogeny}

Now that we have seen how communication works in non-human animals, let us turn to how children start communicating throughout their development. Around their first birthdays, children start communicating ostensively by pointing (Tomasello, 2008). Although at first glance pointing may seem like a simple behavior, it may be used in a number of communicative contexts to convey a fairly 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).

Check the exact timing of this: 9 or 12 months? Big difference. Directly cite experimental sources.

On the classic account, pointing can serve either of two communicative motives: an imperative motive, in which the pointer requests things from some-

one, and a declarative motive, in which the pointer shares their experiences and emotions with someone. This account can be extended upon by distinguishing between declaratives as expressives (sharing attitudes and emotions) and declaratives as informatives (providing information), and by furthermore conceiving 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).

The fact that pointing is a fairly complex communicative act is underscored by the fact that non-human animals are not able to understand pointing in the same way humans are. The hypothesis is 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; that without skills of shared intentionality, infants could only communicate intentionally, but not cooperatively. 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.

Add reference from Tomasello (2008)

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.

Give example

Give this empirical 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. Children acquire the skills for lying around the age of four (Lee, 2013).

Build upon this with stuff from Meibauer (2018)

2.4 What is the function of communication?

{sec:comm:function}

Finally and arguably most importantly for our endeavor, let us have a look at the 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 Seyfarth and Cheney's (2003) review of 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; if their interests do not overlap, or even oppose each other, their communication can be considered to be competitive. 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 commu-

nication for cooperative purposes – i.e., to coordinate their hunting activity. On the other hand, 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. They might for example intimidate each other verbally, 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) and echoed by Dor (2017), the cooperative setting constitutes the 'birthplace' of the unique features of human communication; the competitive use of human-style communication must have emerged later. As Tomasello (2008) writes:

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

Especially the emergence of language could only have occurred in cooperative settings, Tomasello (2008) and Dor (2017) argue. I return to this line of thought in a bit.

Let us now consider pragmatic communication; specifically, how the 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)

Earlier along the evolutionary timeline, an interlocutor's Gricean cooperativeness may very well have coincided with her cooperative intention or disposition. This is of course especially the case with animals communicating (see Section 2.2). However, it is apparent that these two 'dimensions' of cooperativeness need not always coincide. Daniel Dor's (2017) discussion of lying and the evolution of language relates to this distinction between what we might refer to as Gricean cooperativeness and cooperative intent. Let us briefly consider some of the points he makes in this paper before we return to the distinction between Gricean cooperativeness and cooperative intent.

Dor (2017) notes that the distinction between honesty and deception 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. On the other hand, one may consider the honesty of a signal to refer to not its truthfulness, but rather 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. Give example

Dor then outlines four possible communicative options one might choose: "co-operative honesty, harmful honesty, co-operative lying and harmful lying"

(Dor, 2017, p. 45). He argues that while anti-social, exploitative lies – lies with the intention to profit 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. Meibauer (2018) has useful additions to this point: he 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)

Going back to Grice's cooperative principle, note that Dor in his analysis only considers Grice's maxim of Quality ('Try to make your contribution one that is true'). If we extend his distinction to the whole cooperative principle – in other words, if we also incorporate the maxims of Quantity, Relation and Manner – we get the two 'dimensions' of cooperativeness we considered earlier: Gricean cooperativeness, and cooperative intent. These two dimensions combine to give us four communicative options one might choose: Gricean cooperative cooperation, Gricean noncooperative cooperation, Gricean cooperative noncooperation, and Gricean noncooperative noncooperation. For example, lying constitues Gricean noncooperation, because it violates the maxim of Quality. As noted before, lying may be used with competitive intent (i.e. antisocial lying), or cooperative intent (i.e. prosocial lying, such as lying for politeness or white lies).

A small note on language

Let me briefly address the elephant in the room: when considering human communication, human language and its evolution cannot remain unmentioned. In this thesis, I only consider human communication in general, because the emergence and evolution of symbolic communication in the form of language is an entire field of research of its own. I will finish by noting one important thing about the evolution of language as it relates to cooperation and trust. It has been argued that only in cooperative settings could our complex language have emerged at all. This is because for such complexity 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 the idea of 'getting the system going' in Section 2.4.2. For more discussion on the evolution of language, see for example Tomasello (2008) and Dor (2017).

Now, 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 comprehensive 2009 book *Why We Cooperate*.

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 them interchangeably. He defines collaboration as working together for mutual benefit (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 his 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)[2]. He elaborates by noting that 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 brings disadvantages with it when it comes to foraging for food, since the members of the group are competitors in the acquisition of 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 could have been evolutionarily advantageous to be more tolerant and less competitive, which would explain its having evolved. Secondly, Tomasello notes it could be the case that humans went through a process of self-domestication, which eliminated aggressive, predatory or greedy individuals from the group; see Benítez-Burraco et al. (2021) for more on this. 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

[2]Notably, what exactly this selective pressure is, is a missing link in his otherwise very convincing story.

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 can be 'built', so to speak: they provide an environment in which more elaborate collaboration can 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. As before, there is a missing link in this story, in this case it concerns how mutual expectations between individuals arise and eventually become norms. (Tomasello describes it 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 within the collaborative activity. Individuals recognize their dependence on each other for 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 briefly summarize the evolutionary timeline of human cooperation according to Michael Tomasello. At some point, for reasons as of yet unknown to us, foraging for food collaboratively rather than individualistically became beneficial – perhaps even necessary – for humans. During this evolutionary process, some degree of tolerance and trust must have emerged between those 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 as well as a role division for the joint activity. This kind of collaborative activity then constituted the breeding ground for human cooperative communication. These 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 the story, we turn our attention to a crucial aspect of understanding human communication: how it can have persisted despite evolutionary pressures treathening its stability. To this end, let us consider at length a paper by Scott-Phillips (2008), who convincingly brings findings from animal signaling research into the realm of human communication. This will provide a good background for discussing two precursory papers to the argumentative theory of reasoning, by Sperber (and others) in Sections 3.1 and 3.3.

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 as 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[3] 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: if it can be advantageous for me to lie, deceive or mislead someone, then it would evolutionarily make sense for me to do so; yet then it would make evolutionary 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 reliability of animal communication in general. One such 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 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.

However useful in explaining some cases of the reliability of animal communication, the handicap principle is not able to explain all of those cases: often, it is not the case that reliable signals are costly to produce (Scott-Phillips, 2008) . Add example

[3] See Section 2.4.3 for a terminological comment on reliability versus honesty.

Especially in the case of human communication, the handicap principle cannot account for its reliability, since it is in general not costly to produce utterances (Scott-Phillips, 2008). Thus, it remains to be shown how communication can be stable if signals are cost-free.

On the handicap principle, reliable signals are costly to produce, thus ensuring their reliability. An alternative explanation of the reliability of animal communication is the principle of *deterrence*, whereby *un*reliable signals are costly to produce, and consequently signalers are deterred from producing unreliable signals. 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. 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.

possibly explain this more

The 'logic of deterrents' applied to the case of human communication poses the following demands in order for the story about stability to work:

> 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 recognized as unreliable relatively easily, and unreliable signalers incur costs for their unreliability, the reliability of communication is secured through deterrents.

Scott-Phillips goes on to state 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 the first sufficient condition is met in the case of human communication; we will return to this in Section 4.2.

Now, before we consider Sperber's precursory concepts to the argumentative theory of reasoning, it will be good to have a closer look at deception. The 'stability of communication' problem hinges on the assumption that it is advantageous to deceive others. This is intuitively plausible; however, it deserves some extra attention, as this is such a fundamental assumption.

2.4.3 Deception and lying

{sec:deception}

Let us start off this section by clarifying and examining some of the terminology surrounding honesty and dishonesty.

Briefly returning to Scott-Phillips (2008), he uses the term 'reliable' when explicating his story about the stability of communication, rather than 'honest'. As he explains in the paper's introduction, this is a principled choice, to do with a difference between humans and non-human animals. He argues that

one may want to steer clear from anthropomorphically ascribing intentions to animals, and meanings to their behavior. He maintains that thus 'reliability' would be a more neutral term than 'honesty', because it refrains from ascribing intentions and meanings to individuals. However, I find that talking about 'reliable communicators' blurs an important distinction between honest, benevolent communicators and competent communicators. We return to this distinction in Section 3.3.

Deception, lying and persuasion may be defined in a number of different ways. Let us now briefly look to the literature to obtain a working definition of these concepts.

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: Jennifer Mather Saul even dedicates the whole first chapter of her 2012 book to obtaining a viable definition that includes the relevant examples and excludes the irrelevant ones. An intricate discussion of her definition is out of scope for this thesis; let us for now consider a definition of lying due to Williams (2002) that Meibauer (2018) calls 'standard':

> 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)

Meibauer notes – and this also transpires from Saul's (2012) discussion of the definition of lying – that each of the components of this definition can be, and has been, challenged.

Let us now discuss some of the points that Daniel Dor (2017) makes about lying and the stability of communication.

Especially important are the notes he makes regarding what he terms the "paradox of honest signaling" (Dor, 2017, p. 46) – i.e., the theoretical issues plaguing the stability of communication that we discussed in Section 2.4.2. Dor notes that this 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, this story appeals not to receivers evaluating the truthfulness of incoming information, but rather the receivers evaluating 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 appeals to situations in which interests between interlocutors conflict; however, these situations might not be the most pertinent or prevalent kind of communicative situation. According to Dor, 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.

Moreover, (and briefly returning to the utility of communication), another avenue Dor explores is how communication is used. As we will see in Sec-

tion 3.1 and in Section 3.3, Sperber and his colleagues focus a lot on the transmission of information between individuals, in the form of testimony and argumentation. However, Dor argues that 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; Dor writes

> 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 information 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)

Convincingly, Dor goes on to argue that 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. He writes:

> 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)

Lastly, it would be good to 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.[4] 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. For limitations of time and space we will not go into the benefits of gaining social influence. However, I believe the discussion of cooperation in Section 2.4.1 is sufficient for the present purposes.

2.5 Conclusion

Add chapter summary / concluding remarks / transition into Chapter 3

[4] In this, one may see a parallel with Seyfarth and Cheney's (2003) conception of aggressive communication 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 (Mercier and Sperber, 2009; Sperber, 2001; Sperber et al., 2010). 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}

First up, 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

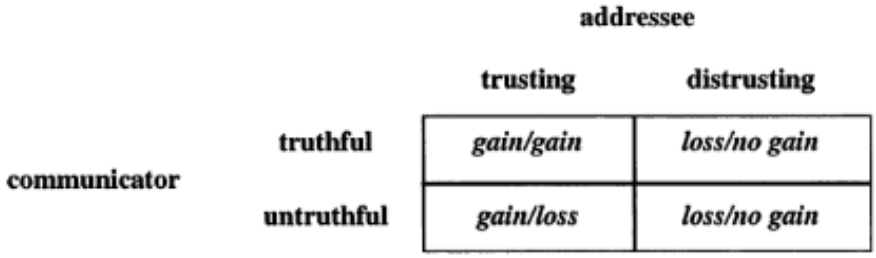
[1] This may very well be the case philosophically or epistemologically speaking, but psychologically speaking, they may be more intertwined than Sperber implies. In Sperber et al. (2010, §3), he and his colleagues elaborate more on his stance, making even stronger claims about how comprehension always 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 Figure 3.1.



{fig:matrix}

Figure 3.1: Payoff matrix for one-off communicative event

Final version: replace this screenshot by TiKZ picture

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)

(Add small summary, concluding remark or segue)

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. They introduce their distinction between intuitive and reflective inferences as part of a massive-modularist framework. This framework warrants some explanation; although I consider the modularity of the mind to be out of scope for this thesis, Mercier and Sperber's dual system theory is best explained and understood through their views on modularity.

Mercier and Sperber are proponents of a massively modular view on the human mind, maintaining that the mind consists of a number of cognitive modules specialized to a specific domain. These modules are autonomous in their function, have distinct evolutionary and developmental histories, and they have characteristic inputs, procedures and outputs. On this massively modular view of the mind, inferences can be performed by many different domain-specific modules. Even inferences that seem to be performed by domain-general module, such as logical inferences, are carried out by domain-specific *metarepresentational modules*: modules that perform inferences on conceptual representations. Because these conceptual representations may belong to any domain, metarepresentational inferences appear to be domain-general. However, Mercier and Sperber state, this domain-generality is indirect and virtual. In their words,

> Metarepresentational modules are as specialized and modular as any other kind of module. It is just that the domain-specific inferences they perform may result in the fixation of beliefs in any domain. (p. 153)

Mercier and Sperber 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 we would refer to as reasoning, or 'reasoning proper', as Mercier and Sperber call it. When it comes to the exact definition of these two categories of inferences however, we unfortunately run into some ontological problems. That is, Mercier and Sperber define intuitive and reflective inferences in two different, seemingly incompatible ways. First, they state that

> intuitive inferences the conclusion of which are the direct output of all inferential modules (including the argumentation module), and reflective inferences the conclusions of which are an indirect output embedded in the direct output the argumentation module (*sic*) (pp. 155–156)

In other words, the *conclusion* of an intuitive inference is the direct output of an inferential module, and the *conclusion* of a reflective inference is an indirect output of the argumentation module. However, later they state that

> Intuitive inferences are the direct output of many different modules. Reflective inferences are an indirect output of one of these modules. (p. 156)

We will return to these problems in Section 4.3. For now, let us briefly pivot to a later discussion of intuitive inference and argument in Mercier and Sperber (2011, §1.1) to further clarify some terminology.

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, they emphasize that their dual systems approach is different from the classical distinction between system 1 and system 2 reasoning. They maintain that system 1 and system 2 operate at the same level, whereas intuitive and reflective inference do not: intuitive inferences are carried out by any module, but reflective inferences result specifically from the argumentation module – more precisely, reflective inferences are an indirect output of the argumentation module (Mercier and Sperber, 2009, p. 156).

Can I assume familiarity with dual-process theories, or do I need to explain them?

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 and moreover, 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 function of reasoning and empirical evidence corroborating these. 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 concerning these.

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 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[2], because they work in tandem to facilitate trust calibration on the side of the receiver.

[2]Slightly confusingly, Sperber et al. call the ostensive-inferential communication that we saw in

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.

Tie this to the discussion of the ostensive-inferential model in Section 2.1

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

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.

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 evolved the means to) exercise *epistemic vigilance* in order to filter incoming information (Sperber et al., 2010).

(5) In order to then get her 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

Doing what? One example would be illustrative

The second prediction is that reasoning shows a confirmation bias.

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

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.

3.5 Conclusion

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: strongly tethering the theory to this cooperative context would make the story as a whole more plausible.

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? Try to pinpoint the exact fallacy

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?

Add in more quotes that point to Sperber's vagueness?

Add an illustrative, intuitive example about this

These different readings of epistemic vigilance are also implicit in the following statement Sperber and colleagues make:

> 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 response paper (2013). 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. Further, he argues that the weak reading of the argument carries with it the assumption that the benefits of epistemic vigilance outweigh its costs.

What is Michaelian's conclusion from this?

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?

In the original 2010 paper, Sperber and colleagues maintain that it is sufficient that communication is *on average* advantageous to both parties:

> The fact that communication is so pervasive despite [the risk of misinformation] suggests that people are able to calibrate their trust well enough to make it advantageous on average to both communicator and audience (Sperber et al., 2010, p. 360)

However, I believe that this does not constitute a full solution to the problem. In order to provide a satisfactory answer, one would need to provide a detailed sketch of costs and benefits.

Does this quote point to them adhering to the 'merely advantageous' reading?

This point is a bit incoherent as of now: more thinking

4.2.4 Honesty or dishonesty as prior

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

Rewrite in academic English

General gist: nice concept, sounds intuitive and convincing enough. But details are lacking, which raises some eyebrows. That's not to say that these details could not be added; but the evolutionary story is not as strong as they take it to be, so they have their work cut out for them. This leads us nicely into the next section.

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

Rewrite in academic English

Details are severely lacking. Evolutionary framework is too narrowly focused. Metatheoretically, it gives bad vibes. None of these issues are fatal flaws (yet), because I don't feel confident to say that details cannot be added, the framework couldn't be expanded. The theory is attractive and intuitively convincing, but internally weak, and requires a lot more effort to fortify it.

5 | Conclusion

{ch:conclusion}

5.1 Further research

{sec:further-research}

One obvious avenue for further research, as already implied in Section 4.1, would be to position the argumentative theory of reasoning within the broader evolutionary context of human cooperation as it transpired from Section 2.4.

5.1.1 What is reasoning? Revisited

{sec:def-scrutiny}

However, much more remains to be said about this very definition. Let us first briefly consider other definitions of reasoning from the literature, and then place some question marks around the position of Mercier & Sperber's definition of reasoning within their argumentative theory.

The definition of reasoning given and used by Mercier & Sperber is reminiscent of definitions of argumentation, in particular in its use of the terms 'premise' and 'conclusion'. In a 2001 paper, which can be considered to be a precursory work to Mercier and Sperber (2011), Dan Sperber uses the following definition of argumentation:

> the defense of some conclusion by appeal to a set of premises that provide support for it (p. 401)

and somewhat less precisely, Mercier and Sperber (2011) define arguments as

> representations of relationships between premises and conclusions (p. 58)

Comparing these definitions of argumentation and of reasoning raises a question: what is reasoning, if not internalized argumentation? Or, in a similar vein, what is argumentation, if not externalized reasoning? And, if this is the case, does this not render the argumentative theory of reasoning void? For then the theory would state that the main function of internalized argumentation is argumentative.

5.1.2 Motivations and dispositions of interlocutors

Throughout their 2011 article, Mercier & Sperber allude to the dispositions of interlocutors in argumentative settings (emphasis in quotes added):

'Dispositions' is not meant as a technical term, and I don't think it is; is it?

> This experiment illustrates the more general finding stemming from this literature that, *when they are **motivated**, participants are able to use reasoning to evaluate arguments accurately* (p. 61)

> Most participants are **willing** to change their mind only once they have been thoroughly convinced that their initial answer was wrong (p. 63)

> this [experimental finding] should not be interpreted as revealing a lack of ability but only a lack of **motivation**. When participants **want** to prove a conclusion wrong, they will find ways to falsify it. (p. 65)

> people are good at assessing arguments and are quite able to do so in an unbiased way, **provided they have no particular axe to grind**. In group reasoning experiments where participants **share an interest in discovering the right answer**, it has been shown that *truth wins* (p. 72)

This reference to the motivations and disposition of interlocutors opens up some questions as to the specifics of the 'argumentative setting' that Mercier & Sperber mention multiple times throughout the paper. It seems that the disposition of the interlocutors going into an argument plays an important role in Mercier & Sperber's account of argumentation, yet they do not expand on this. How plausible is the assumption that people engaging in argumentation have a 'common interest in the truth', as Mercier & Sperber call it? And what happens (or what would happen) when interlocutors do *not* share this interest?

> When people are motivated to reason, they do a better job at accepting only sound arguments, which is quite generally to their advantage (p. 96)

To do:

- I'm pretty sure this criticism is about something different than 'motivated reasoning', but check this
- Define what an argumentative setting is, according to Mercier & Sperber (close-read Mercier and Sperber (2011) for this)
- Possibly find some empirical work on arguers' dispositions

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.

Baedke, J. (2021). "What's wrong with evolutionary causation?" In: *Acta Biotheoretica* 69.1, pp. 79–89.

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.

Buss, D. M. (2015). *Evolutionary psychology: The new science of the mind*. Fifth. Routledge.

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.

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.
Day, R. L., K. N. Laland, and F. J. Odling-Smee (2003). "Rethinking adaptation: the niche-construction perspective". In: *Perspectives in biology and medicine* 46.1, pp. 80–95.
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. 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.
Grice, H. P. (1975). "Logic and conversation". In: *Speech acts*. Brill, pp. 41–58.
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.
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.
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.
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.
Mayr, E. (1961). "Cause and effect in biology". In: *Science* 134.3489, pp. 1501–1506.
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.
Popper, K. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge.
Quine, W. V. O. (1960). *Word and object*.
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.
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.
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. (2008). *Origins of human communication*. MIT Press. DOI: 10.7551/mitpress/7551.001.0001.
— (2009). *Why we cooperate*. MIT Press.
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.
Wason, P. C. (1968). "Reasoning about a rule". In: *Quarterly journal of experimental psychology* 20.3, pp. 273–281.
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.