ABSTRACT

Title of Dissertation:	MINDREADING FOR COOPERATION: A MODERATELY MINIMALIST APPROACH
	Julius Schönherr, Doctor of Philosophy, 2019
Dissertation directed by:	Professor Peter Carruthers, Department of Philosophy

This dissertation puts forth a series of arguments about the extent to which human cooperative interaction is fundamentally shaped by *mindreading*; i.e. the capability to reason about the psychological causes (e.g. intentions, beliefs, goals) of behavior.

The introduction to this dissertation discusses the broad philosophical underpinnings that lay the foundations for more specific philosophical issues under discussion in subsequent chapters.

In chapter two, I argue that a thorough interpretation of the relevant empirical evidence suggests that mindreading is fast, effortlessly deployed, and operative sub-personally. For this reason, mindreading is principally well-suited to enable most everyday cooperative interactions. In the appendix, I (in collaboration with Evan Westra[1]) elaborate on this picture, arguing that the cognitive mechanisms operative in

[1] Equal contribution by both authors.

social interactions are, in all relevant respects, similar to those operative in non-interactive situations.

While chapter two and the appendix defend the idea that the cognitive faculties responsible for mindreading are fit to enable cooperative interactions, chapters three and four take this perspective for granted and discusses whether human cooperation is crucially dependent on a form of *reciprocal* attribution of mental states that is often labeled *common knowledge*.

In chapter three of this dissertation I address, and reject, the oft defended idea that truly performing an action together with others requires that all parties commonly know their intended goals. I argue that this view is fundamentally mistaken. Successfully acting together with others often *requires* not knowing these goals.

Chapter four explores reciprocal belief attribution in the context of coordination problems. Humans often coordinate their actions by replicating successful past choices; they reason based on *precedent*. Philosophers have often claimed that solving coordination problems by relying on precedent presupposes common knowledge that all parties rely on precedent in trying to coordinate their actions. Chapter four points out that this assumption is erroneous: Coordinating behavior on the basis of precedent is broadly incompatible with any higher-order knowledge (or beliefs) about the other agents' choices.

MINDREADING FOR COOPERATION: A MODERATELY MINIMALIST APPROACH

by

Julius Schönherr

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Dan Moller
Professor Eric Pacuit
Professor Christopher Morris
Professor Brian Kogelmann
Professor Melanie Killen

Dedication

For Utsch

Acknowledgements

I am deeply indebted to a great many people who have supported me throughout my time writing this dissertation, throughout my academic career more generally, and, ultimately, throughout my life as a whole. My greatest debt is to the members of my immediate family; my mother Martina Schönherr and my brother Max Schönherr. Martina has always been a source of love and encouragement. Max is my greatest source of inspiration and my best friend. My other great debt is to my supervisor, Peter Carruthers, who is an incredible philosopher and whose work I deeply admire. Peter has been unbelievably supportive at every stage of my dissertation project. He has diligently commented on and reviewed my work countless times. He has given me invaluable career advice; and he is an unbelievably kind and generous person. I also want to thank Dan Moller who has encouraged me to be bold and think independently. Many thanks are also owed to Eric Pacuit, Brian Kogelmann, Christopher Morris, and Thomas Schmidt who have played important roles in my development as a scholar. I also want to thank Evan Westra with whom I've had the pleasure of collaborating on the appendix of this dissertation project. I am fortunate to have received support and comments on my work from a number of other faculty, friends, and colleagues: Andrew Fyfe, Javier Gomez-Lavin, Quinn Harr, Aleks Knoks, Moonyoung Song, Javiera Perez Gomez, Hans Bernhard Schmid, and Aiden Woodcock. I am also grateful for comments from anonymous reviewers and editors from the following journals: Philosophical Psychology, Review of Philosophy and Psychology, and British Journal of the Philosophy of Science.

Table of Contents

Chapter 1: Foundations and Motivations

1. *Introduction*

This dissertation is comprised of three core chapters and one appendix. All but chapter four have published in journals.[2] Each chapter discusses one aspect of the importance of mindreading – i.e. the capacity to reason about the mental states of other agents[3] for the purposes of predicting and interpreting behavior – in cooperative contexts. In chapter two, I argue that mindreading is regularly and pervasively deployed when interacting with others. In this sense, as I, in collaboration with Evan Westra[4], demonstrate in the appendix, there is no special problem of mindreading in the context of social interaction; most cognitive systems designed for mindreading work the same in interactive as well as non-interactive contexts. Chapters three and four turn to the importance of *common knowledge* – a type of *reciprocal* mental state attribution – in two cooperative contexts: coordination games and joint action. In both chapters, I conclude that the importance common knowledge has been thought to play in these contexts has been overstated in prior philosophical work. Successful coordination and joint action do not, at least not always, require all party's intentions to be commonly known between them.

---

[2] Chapter two: Schönherr, J. (2017). What's so Special About Interaction in Social Cognition?. *Review of Philosophy and Psychology*, *8*(2), 181-198.
Chapter three: Schönherr, J. (2018). Lucky joint action. *Philosophical Psychology*, 32:1, 123-142.
Appendix: Schönherr, J., and Westra, E. (2019). Beyond 'Interaction': How to Understand Social Effects on Social Cognition. *The British Journal for the Philosophy of Science,* 70(1), 27-52.
[3] Arguably, mindreading can also be turned onto the self to interpret one's own mental states. For the purposes of this dissertation we can ignore this possibility.
[4] Equal contributions by both authors

The four parts of this dissertation are independent of one another and can be read in any order. However, these chapters share a common approach, and can be read as part of a broader project. This project starts with the conviction that cooperation is an essential feature of human life and that the skilled deployment of mindreading is crucial for enabling and sustaining successful cooperation. However, provided that mindreading is fundamentally involved in solving a range of cooperative tasks, this dissertation contributes to an exploration of the precise extent to which mindreading is required to satisfy this function.

2. Rescuing Tommy's cat

> **Rescue Mission.** "Tommy's cat Cuddles escaped a minute ago. I just saw her in the backyard. I'm sure she's still close by. Would you help me find her?", says Ian. Mia, who's quite fond cats, agrees; "yes, of course! What does she look like?". "She's ginger, but the tips of her ears are white; quite cute, you'll see", says Ian. And off they go to capture the cat. As they round the corner, Ian sees a ginger cat with white ear tips; only that it is clearly not Cuddles. He turns to Mia and notes "Oh, that's Loge, not Cuddles, Let's keep looking". They turn around and see Cuddles sitting on the sidewalk in front of Ian's house calmly chewing on the remains of a mouse. Ian looks at Mia and then veers off in the direction of his front door. He opens the door and then nods in the direction of Mia who now tiptoes behind the cat. She makes a loud noise. The poor cat, startled by the

sudden eruption, steams off into the house. Success! – the cat is back in the house.

Although this vignette does not explicitly state that Ian and Mia attribute mental states to one another, a discussion of the psychological processes operative in their rescue mission will, almost inevitably, make reference to such mental state attribution. Mia's curt answer "yes, of course" indicates to Ian that she *wants* to help and that she will, thereafter, *try* to find the cat with him. Furthermore, the reason why Ian tells Mia, upon seeing the first cat, that it is Loge, not Cuddles, seems readily explained by him noticing that Mia *falsely believes* that it is Cuddles she is looking at. When Ian veers off to open his front door, he counts on Mia to understand his plan; his intention for the both of them, that is. A final, admittedly controversial, feature of their rescue mission may be that they both not only know that they are looking for the cat, but also that they both know that they both know that they are looking for the cat, that they both know that they both know that they both know this, and so on *ad inf.*

In this dissertation, I will answer three specific questions about the nature of mindreading in cooperative contexts akin to 'Rescue Mission'. I will argue that mindreading-involving accounts of such cooperative interactions are basically right (chapter 2, appendix), but I will also point to some of the limits of these accounts. Extensive iterative knowledge of each cooperator's intentions (e.g. that they both know that they both know these intentions) is not always required to cooperate successfully.

Before delving into the thick of things, addressing these specific questions, let me use this introduction to give an aerial depiction of theories and assumptions relevant to my discussion in subsequent chapters of this dissertation. Let's proceed as follows: In section 3, I will canvass an important line of research according to which the extensive use of mindreading for cooperative purposes is one fundamental keystone in the evolution of distinctively human culture. In section 4., I will turn to *reciprocal* attributions of mental states and explain why philosophers have traditionally thought that various forms of cooperation (e.g. joint action, coordination, convention) must involve a particular form of reciprocal attribution of mental states often referred to as '*common knowledge*'.

3. *Mindreading, cooperation, and human evolution*

"Humans", Joseph Henrich contends, are "a puzzling primate" (Henrich 2016, vii). On the one hand, human life is impressively different from the life of other primates. Humans, for instance, live in more environments than any other terrestrial mammal, use more complex technologies, live in intricate social arrangements, and follow a range of moral and social norms (Henrich 2016, ix). But while illustrating human distinctness by pointing to impressive achievements rooted in cooperative culture is straight-forward, identifying the cognitive capacities responsible for these human-specific achievements is less obvious; and it is even less obvious which among all relevant capacities are the fundamental ones.

In this introduction, I cannot, of course, hope to provide a comprehensive survey of all research in philosophy, cognitive science and evolutionary anthropology that

may help understand the cognitive basis for distinctively human social life. Rather, I will describe, in reasonable detail, an emerging consensus that it is the human expertise and motivation to use mindreading for cooperative purposes that lays a partial, but important, foundation for human-specific cooperative life. Talk of a "foundation" is meant to indicate an evolutionary perspective. The cooperative use of mindreading is what *set human sociality apart* from the lives of other primates *in the ancestral environment*. There are several arguments to support this conjecture.

First, in a direct experimental comparison between cognitive capacities of children and chimpanzees, children's far outstrip chimpanzees' capacities only along social cognitive dimension. Second, there is little evidence that apes are *motivated* to use their constrained capacity for mindreading for cooperative purposes. Thirdly, this is true even in the case of cooperative hunting, arguably the most plausible contender case for mindreading-enabled chimpanzee cooperation. Let me delineate each line of argument in turn.

In a series of experiments, Herrmann *et al.* (2007, 2010) subjected preschool children and chimpanzees to a battery of experiments, testing each group's general intelligence, spatial, and social cognitive capacities. In these experiments, humans and primates, had to track moving objects, discriminate quantities of objects, engage in causal reasoning about hidden rewards, interpret communicative clues about a reward's location, follow an actor's gaze, emulate an actor in solving a task. Hermann and colleagues found that primates and humans differed significantly only along the social cognitive dimensions (i.e. gaze following, and social learning). These

differences are a matter of degree; although, in these experiments, primates were able to learn from others and follow others' gaze, the difference between primates and children was pronounced.

Now, being able to follow gaze, or learn from others does not, by itself, demonstrate an understanding of mental states; i.e. being able to follow someone's gaze does not imply understanding that others *attend* to things; and being able to learn from others does not imply an understanding that others *intend* to demonstrate something. However, many (e.g. Hare *et al*. 2001, 2006; Melis *et al*. 2006; Buttelmann *et al*. 2007; Call and Tomasello 2008; Kaminski *et al*. 2008) believe, on independent grounds, that primates and other animals are capable of understanding a range of mental states (e.g. know what others know, perceive, intend).[5] It is, thus, a reasonable conjecture that these performance differences in social cognitive tasks are explained, at least in part, by differences in an underlying social cognitive capacity to attribute mental states to others.

While the difference between humans and primates with regard to their understanding of mental states such as *seeing, knowing,* and *intending* is a matter of degree, the differences with regard to their understanding of *belief* may turn out to be categorical. Neurotypical humans, on the one hand, seem to attribute beliefs[6] to others with ease and early on. This much is uncontroversial. Controversies revolve

[5] For a dissenting voice, please consult Penn and Povinelli (2013) who nevertheless concede that "most would probably agree […] that nonhuman animals 'understand some psychological states in others – the only question is which ones and to what extent'" (Penn and Povinelli 2013, 62f).
[6] Belief understanding is typically measured in terms understanding *false* belief. As philosophers (Dennett 1978; Premack and Woodruff 1978) have pointed out, in order to show that an individual possesses the concept 'belief' one must show that this individual understands the representational nature of this mental state.

around the precise developmental trajectory of such understanding. Until the early 2000s, there was a broad consensus that children start understanding beliefs at around age 4 (Wellman *et al.* 2001). However, more recent research, initiated by Onishi *et al.*'s (2005) seminal study[7], uses non-verbal measures (e.g. looking time, active helping) and finds evidence for belief understanding in infants between 6 and 36 months (e.g. Wang and Leslie 2016; Senju *et. al.* 2011, Kovács *et al.* 2010; Onishi and Baillargeon 2005; Southgate *et al.* 2010; Southgate and Vernetti 2014). Now, while there are over "30 reports spanning 11 different methods, providing convergent evidence for false-belief understanding in children ages 6–36 months" (Baillargeon *et al.* 2018), some of these findings could not be replicated (e.g. Southgate *et al.* 2007), some have been replicated successfully (e.g. Wang and Leslie 2016; Senju *et. al.* 2011), and many experiments have simply not been attempted to be replicated. For a detailed discussion of these replications problems consult Baillargeon *et al.* (2018) who hypothesize that "procedural differences between studies may explain failures to replicate" (Baillargeon *et al.* 2018, 112).

Alternative interpretations of experimental results from implicit false belief tasks, however, have been offered. Butterfill and Apperly (2009, 2013) distinguish between two distinct cognitive capacities: on the one hand, a capacity to understand beliefs "as such" (Butterfill and Apperly 2013,607); on the other hand, a cognitively limited capacity "track" others' beliefs (Butterfill and Apperly 2013). Non-verbal false-belief tasks, it is argued, provide evidence for the latter, but not the former, capacity

[7] In this study, Onishi *et al.* used a looking time paradigm to show that 15 months olds predict an actor's behavior taking into account the actor's false belief about the location of an object.

(Butterfill and Apperly 2013, 620). Heyes (e.g. 2014) argues that infants' performance in non-verbal false belief tasks can be explained by "low level novelt[ies]" (Heyes 2014, 647) (e.g. the relative novelty of a green, as opposed to red, boxes). The best experimental evidence, Heyes contends, is compatible with both, cognitively rich and sparse, interpretations.

Non-human animals, on the other hand, mostly fail to demonstrate belief understanding (e.g. Kaminski *et al*. 2008; Krachun *et al*. 2009, Call and Tomasello 2008). Challenging this consensus, two recent studies (Krupenye *et al.* 2016; Buttelmann *et al.* 2017) find evidence that great apes distinguish true from false beliefs in a looking time paradigm (Krupenye *et al.* 2016) and an interactive helping task (Buttelmann *et al.* 2017). More research is required to reach a conclusive verdict on belief understanding in non-human animals.

Let's turn to the second of the above-mentioned arguments. Evidence suggests that, in the context of cooperative interaction, primates make *very limited use* of their capacity engage in mindreading. The default case for this claim comes from experimental research indicating that great apes perform poorly in cooperative tasks (e.g. Melis *et al*. 2006; Silk *et al*. 2005; Hamann *et al*. 2011, Hare and Tomasello 2004). In each of these studies it is argued that, although apes do cooperate, their motivation is highly *limited*. For instance, Melis *et al*. (2006) found that chimpanzee cooperation breaks down when subsequent rewards require food sharing. Chimpanzees cooperate more when the rewards are presented in separate piles, one for each collaborator, or when the costs of cooperating are low. Hamann *et al.* (2011)

found that, while chimpanzees do, at times, make food available to others, such food sharing does not correlate with prior collaboration on a task. Lastly, Hare and Tomasello (2004) found evidence across a range of experiments that chimpanzees perform better in competitive as opposed to cooperative settings. All this leads to the assessment that great apes have a "Machiavellian intelligence. […] [G]reat ape social cognition seems built for outcompeting others by outsmarting them" (Tomasello 2016, 22).

Humans, on the other hand, show strong prosocial tendencies early on and are highly motivated to use their cognitive capacities to facilitate cooperation. Buttelmann *et al*. (2009) found that children actively help an adult taking into account the adult's false beliefs. Children between 9 and 15 months expect others to help those in need (Köster *et al.* 2016) and they avoid agents with harmful intentions (Vondervoort *et al.* 2017). In the context of mutualistic cooperative activities, three-year-olds share the rewards of their collaboration equitably, even when the chance to selfishly monopolize rewards is available (Warneken *et al.* 2011). Twelve-month old children will spontaneously provide helpful information to a partner by pointing to a desired object (Liszkowski *et al.* 2006). This is but a snapshot of the existing literature on infant cooperation, but it shall suffice to demonstrate that, while chimpanzees and other great apes use their mindreading skills mostly in competitive contexts, humans collaborate freely, effortlessly, and early on.

Now, although this experimental research suggests that great apes use their mindreading skills largely for competition, some have argued that these results are

readily explained by the artificial experimental set-up of the relevant experiments. For this reason, some argue, these experiments fail to generate viable hypotheses about cognitive capacities in non-human animals. More ecologically valid research is required to justify such conclusions. More particularly, in a series of studies, Boesch and colleagues (e.g. Boesch 1994; Boesch and Boesch-Achermann 2000) have argued that chimpanzee hunts of monkeys in the Tai forest are deeply cooperative and mentalistic. Allegedly, proper theorizing about these hunts shows that chimpanzees are capable of understanding shared goals and complex social roles in a cooperative task.

Over the years, this rich cognitive characterization has been subjected to extensive criticism. First, and most importantly, this interpretation is simply at odds with the experimental findings summarized above. Secondly, alternative, less complex, but experimentally substantiated explanations have been offered. Tomasello *et al.* (2005, 2011) proposes that, in those hunts, a single chimpanzee begins the hunt on its own. Subsequently, others occupy the "most opportune spatial position still available at any given moment in the emerging hunt" (Tomasello 2011, 8). Importantly, according to this interpretation, each individual simply tries to maximize their personal expected payoff not representing the hunt either as a collaborative effort, nor understanding the hunt as an enterprise that requires specific roles to be filled. This interpretation is experimentally bolstered by Bullinger *et al.* (2011) who find that chimpanzees will use this simple "leader-follower strategy" in controlled experimental set-ups.

Summarizing this discussion, there is a long line of experimental research investigating both humans' and non-humans' capability and motivation to use mindreading in cooperative contexts. Overall, this research suggests that humans, even in infancy, are highly motivated and skilled to use mindreading in an effort to cooperate with others. Non-human animals' skill and motivation to use mindreading to cooperate, on the other hand, is highly limited.

Over the past 15 years, some philosophers and psychologists – commonly labeled "enactivists" – have challenged the idea that cooperative interactions such as hunting, walking together, or baking a cake together need to be explained and sustained through the exercise of mindreading. Mindreading, the worry has it, is a slow and effortful, mental process. For this reason, mindreading could not possibly be the cognitive basis for fast, online cooperative interactions such as group hunts. In chapter two of this dissertation, I will argue that enactivists are simply wrong about this: Mindreading is often fast, spontaneous and does not always require cognitive effort. In the appendix, I, in collaboration with Evan Westra, elaborate on this picture, arguing that the same cognitive mechanisms are operative in interactive as well as non-interactive (e.g. observational) situations.

4. Cooperation and common knowledge

Knowing what others want, know, believe and feel can benefit one's attempts to cooperate with others; this is, of course, a truism. The goal of the previous section was to show just how fundamental mindreading might be for human cooperation. In the previous section, we focused exclusively on first-order mental state attributions.

In the present section, we will expand our discussion and focus on the role of *reciprocal* mindreading (i.e. common knowledge) for cooperative interaction.

Broadly sticking to our hunting paradigm, let's start with an intuitive vignette to introduce the idea of reciprocal mindreading:

> **Failed Hunt.** You and I are hunters and we are given the opportunity to hunt a large stag, too nimble and fierce to hunt individually. Any attempt to hunt down the stag single-handedly would in fact be quite dangerous. Together, however, we stand a good chance of success. Initially, we both think that we each endorse this plan. Suppose however that, for some reason or other, you think that I think that you have decided not to play your part. In this case, you are led to believe that I won't play my part; after all, you know that I don't want to be the sole hunter. For this reason, you predict that you'd be by yourself hunting this game. As a result, you decide not to participate which makes me the sole hunter. The hunt fails. The stag escapes.

In situations such as this one (standardly labeled "Stag hunt"[8]), false higher-order beliefs (beliefs about what the other hunter believes) can stifle our cooperative

[8] To be more precise, 'stag hunt' refers to a game with two Nash equilibria; one strict equilibrium, and one non-strict equilibrium. Such games can be represented by the following payoff matrix:

		Player 1 X	Player 1 Y
Player 2	X	4,4	1,0
	Y	0,1	1,1

efforts. Other times, reciprocal expectations help stabilize our cooperative endeavors. To see this, consider another example:

> **Successful Hunt.** As before, you and I are hunters and we are given the opportunity to hunt a large stag, too nimble and fierce to hunt individually. Any attempt to hunt down the stag single handedly would in fact be quite dangerous. Together, however, we stand a good chance of success. If any of us gave up their intention to participate in the hunt this could potentially introduce danger to the lonely hunter who's left behind. Initially, let's assume, I am uncertain about whether or not you will participate in the hunt. Suppose, however, that I later learn that you believe that I will participate in the hunt, which is why it is reasonable for me to suppose that you won't bail. Given this belief, I, in turn, am motivated to participate.

My initial doubts about you playing your part are dispersed by my thinking that you think that I will cooperate; i.e. these beliefs have assured me that you won't bail.

The idea that the presence (or absence) of these reciprocal beliefs can support (or impede) successful cooperation has played a crucial role in a range of theories; empirical, game theoretic, and conceptual. I will introduce each of these theories; but let me first give a little more general characterization of reciprocal mindreading in terms of *common knowledge*.

"X" indicates the cooperative move. "Y" indicates that defection. These games are called trust games, simply because the only reason to defect could be a prediction that the other player will defect.

In the above examples, 'Failed Hunt' and 'Successful Hunt', we've made explicit how second-order mental state attribution can be relevant to cooperation. We can, however, construct similar arguments for third-order, forth-order, and ultimately *n*th-order beliefs. Consider a slightly amended version of 'Failed Hunt' containing a third-order belief: Suppose I believe that you think that I think that you won't participate in the hunt. In this case, I will be led to believe that you think that I won't participate, which, in turn, rationalizes my prediction that you won't participate. For this reason, given this belief, I ought not to participate. Such arguments, it is easy to see, seem to generalize to *n*th level beliefs which is why successful cooperation, at least in scenarios akin to our hunting vignettes, seem to require an infinite cascade of higher-level beliefs about the all participants' strategy choices; at least this seems tempting to say. This phenomenon has been labeled *common knowledge.* To cut a long story short, a proposition is common knowledge between a group of agents "just in case they all know it, they all know that they all know it, they all know that they all know that they all know it, and so on" (Lederman 2017, 2); and it is this kind of knowledge that seems to play an enabling role for typical cases of cooperative interaction such as the ones introduces above.

Now, characterizing common knowledge in terms of such infinitely nested beliefs is really just that; – a *characterization*, not a definition. Common knowledge merely *entails* such nested knowledge; this is not to say that this nested structure *constitutes* common knowledge.[9] To see the distinction between constitution and entailment,

[9] E.g. Harman (1977), Heal (1978); Milgrom (1981); Clark and Marshall (1981); Mertens and Zamir (1985); Barwise (1988); Lismont and Mongin (2003) give alternative definitions of common knowledge in terms of the publicity of an event, recursive definitions, or probabilistic formulations.

consider David Lewis' work (see Cubitt and Sugden 2003, 185; Lewis 1969) who famously argued that common knowledge is *defined* in terms of symmetrically positioned reasoning on the part of various agents. Such symmetry is then said to *entail* nested knowledge.

This much shall suffice to introduce the idea of common knowledge. Above, I indicated that the idea of reciprocal mindreading (i.e. common knowledge) has been applied to empirical, conceptual, as well as game theoretic questions related to cooperation. In what follows, I will provide a brief survey of these applications. Let's start with empirical applications.

In the previous section, I reviewed a range of empirical findings pointing to the idea that humans' use of mindreading in cooperative situations is at least one fundamental ingredient in specifically human cultural evolution. Beyond what was said above, Tomasello (2005, 2011) hypothesizes that human cooperation is distinct from great ape cooperation in that humans can form *joint goals* which are defined in terms of reciprocal mindreading: "For you and me to form a joint goal (or joint intention) to pursue a stag together, (1) I must have the goal to capture the stag together with you; (2) you must have the goal to capture the stag together with me; and , crucially, (3) we must have mutual knowledge, or common ground, that we both know each other's goal" (Tomasello 2011, 36).

Others have elaborated on this idea arguing that joint *attention* similarly requires such a "common ground" (e.g. Moll and Meltzoff 2011). While researchers such as Tomasello and Moll have *alluded* to cognitive structures involving reciprocal

mindreading in describing human cooperation, they have not provided a more thoroughgoing analysis of this phenomenon. Let's, therefore, move on to appeals to common knowledge in game theoretic models of cooperation.

Game theorists have traditionally been interested in the notion of common knowledge, in part because strategic solutions (i.e. solutions that are not based on luck) to various cooperatives games seem to require common knowledge of each player's strategy choice. Let me illustrate this idea with the following depiction of the famous "coordinated attack" vignette:

> **Coordinated Attack**. "Two divisions of an army are camped on separate hilltops overlooking a valley. In the valley awaits the enemy. If both divisions attack the enemy simultaneously they will win the battle, while if only one division attacks it will suffer a catastrophic defeat. Each of the generals commanding these hilltop divisions wants to avoid a catastrophic defeat: neither of them will attack unless he believes that the general commanding the other division will attack with him. During the night a thick fog descends over the hilltops; the only way the generals can communicate is by sending a messenger through the enemy camp." (Lederman 2018, 921)[10]

It can be shown that rational generals, who have common knowledge that they are rational, will only attack after infinitely many messages have been exchanged. If we then think of each of these messages as a ground for a belief concerning the

[10] The original rendition of coordinated attack goes back to Gray (1978, 465) and was elaborated on in Fagin (1995, 190-1)

respective other's plan, then the generals will only successfully coordinate their efforts given that they commonly know that they plan to attack; i.e. after infinitely many messages have been exchanged. Cooperation in these scenarios, it is argued, requires common knowledge; "almost common knowledge" (Rubinstein 1989) is insufficient to make attacking rational.

For a range of reasons, many philosophers and game theorists have found these conclusions unsatisfying. First, in experimental settings people tend to coordinate successfully on the attack strategy after finitely many messages have been sent (e.g. Camerer 2003; Kneeland 2012). Others (e.g. Binmore *et al.* 2001) argue that thought experiments such as 'Coordinated Attack' fail in important ways to resemble real life cooperation problems, which is why such thought experiments also fail to generate insights about real-life cooperation. Furthermore, Lederman (2018) argues that even ideally rational generals can coordinate their attack after sending finitely many messages, provided that their rationality is not common knowledge.

Recently, however, common knowledge requirements in game theory have been subjected to more radical criticism. Lederman (2017) argues that ideal agents can never have common knowledge. In chapter four of this dissertation, I add to this more radical line of criticism: sometimes cooperation can be achieved by replicating past cooperative success – i.e. we can use precedent to coordinate our actions. Precedent-based solutions, however, are broadly incompatible with *any* higher-order beliefs about other agents' respective strategy choices. The argument, put succinctly, is that

higher-order beliefs function as powerful epistemic defeaters for precedent-based inferences.

These game-theoretic as well as empirical approaches reviewed above are united in highlighting (or questioning) the importance of reciprocal mindreading (e.g. common knowledge) for *successful* cooperation. Philosophers interested in joint action and convention, on the other hand, are interested in reciprocal mindreading (i.e. common knowledge) for the purpose of explicating *what it means* to act jointly or to follow a convention. Let me address the case of convention first and, thereafter, proceed to discussion 'joint action'.

The concept of common knowledge has been invoked in defining conventional behavior. Driving on the right side of the road is a convention, at least in certain parts of the world. What distinguishes such conventional behavior from other, superficially similar, ways of acting (e.g. everyone's driving on the right side simply by happenstance)? This question was famously addressed by David Lewis in his landmark work *Convention* (1969). Lewis' definition of a convention is complex and controversial. For some conduct to count as a convention it must, among other things, be a behavioral regularity – most people drive on right side –, there have to be viable alternatives available – people could be driving on the left –, and all participants (or a relevant subset) must expect all others to act in accordance with this regularity.[11] Furthermore, Lewis argues that these conditions must also be *commonly known* by all participants in the convention. If, for instance, everyone were to expect everyone else

[11] Both requirements are contentious. Gilbert (1981) argues that *de facto* compliance is required. Furthermore, Bicchieri (2006, 34ff) argues that not *all* members of a group have to comply. What is required is that *enough* people comply.

to drive on the right side, but also thought that everyone acted this way for merely frivolous reasons, then, arguably, we would not want to call this behavior conventional. Over the years, many have criticized Lewis' definition. Burge (1975) argues that, although it is true that conventional actions must have alternatives, these need not be commonly known. Gilbert (1981) argues that conventions don't require extensive *de facto* compliance, but, rather, a commitment to comply. Miller (2001) argues that not all alternatives must be equally as good. However, despite these criticisms, each of these authors has invoked common knowledge in their own treatment of conventions. Gilbert, for instance, argues that the relevant commitments need to be common knowledge, and Burge merely argues that *some* (but not all) features of a convention (e.g. the fact that conventions require alternatives) need not be common knowledge. Thus, although the precise role common knowledge plays in defining conventions is controversial, the fact that common knowledge has *some role to play* is not.

Let's finally talk about the role common knowledge has been thought to play in defining 'joint action'. This term, philosophers (e.g., Kutz 2000, 5, Miller 2001; Gilbert 2003; Bratman 1993, 99) contend, refers to a basic way of acting together that is distinguished from mere parallel action. To see the difference between both ways of acting, contrast the case of two strangers walking down Fifth Avenue next to one another, each intending not to run into the other, with the case of two friends walking down Fifth Avenue *together*. Joint action theory sets out to explicate the distinguishing features of such examples. Philosophical orthodoxy has it that one such distinguishing feature is a piece of common knowledge that they both intend to walk

with the respective other person (e.g., Bratman 2013; Miller 2001; Tuomela 2005). Thus, one alleged difference between strangers walking next to each other and friends walking together is that the friends have common knowledge that they are walking together with the respective other friend. In chapter three of this dissertation I argue against this requirement. More specifically, I argue that (a) acting jointly does not require that the participants commonly know their respective intentions, and, furthermore, that (b) acting jointly, at times, depends on the fact that each participant's intentions are not commonly known among them.

This concludes our introduction. We've reviewed a variety of ways – empirical, game theoretic, and conceptual – in which mindreading has been invoked to explain and define ways in which humans can successfully cooperate. Throughout this introduction, we have reviewed theories stressing that mindreading (first order, and reciprocal) is *fundamental* for human cooperation. The individual chapters of this dissertation, to which I have alluded at various points in this introduction, serve to explore targeted questions about the extent to which mindreading may prove to be fundamental in cooperative contexts. The emerging picture is that, while first-order mental state attributions are a fundamental, and experimentally demonstratable, part of human social life, the role common knowledge has been though to play in defining 'joint action' and in solving coordination games has been overstated.

Chapter 2: What's so Special About Interaction in Social Cognition?

1. *Introduction*

For now over 15 years, some researchers in neuroscience, cognitive science, and in philosophy have advocated and defended the idea that our ability to successfully interact with other agents and our ability to understand other agents when we observe them call for different "social cognitive" explanations. However, there is as of now no consensus about how best to describe the "social cognitive" rift that separates interactive and observational contexts. More particularly, on the one hand, researchers disagree about the *kinds* of cognitive mechanisms (e.g. word learning, gaze following, belief-desire attribution, attention allocation) that are recruited differentially in both contexts. On the other hand, there has been widespread disagreement about the *extent* to which cognitive mechanisms are recruited differentially. In this introductory section, I will give an overview over four prominent proposals from the recent literature. Thereafter, I will focus my discussion on one particular interpretation according to which interactive and observational contexts are *categorically* distinct with respect to *belief-desire* attribution.

A number of researchers have construed the interaction–observation divide as a matter of degree. Most prominently, Leonard Schilbach and colleagues' extensive research (e.g. Schilbach 2014; Schilbach *et al*. 2010, 2013A) concerning the

relevance of interaction with regard to lower-level social cues such as mutual gaze, joint attention and socially relevant facial expressions has shown that there are distinct neural activation profiles associated with interactive and observational contexts. Firstly, they use interactive eye tracking[12] to show that self-directed facial expressions lead to "a differential increase of neural activity in the ventral portion of medial prefrontal cortex and the (superficial) amygdala, other-directed facial expressions resulted in a differential recruitment of medial and lateral parietal cortex" (Schilbach *et al*. 2013A, 400). Thus, self-directed facial expressions are associated with "emotional and evaluative processing" (Schilbach *et al.* 2006, 2013B). Secondly, when jointly attending to an object the medial prefrontal cortex, and the posterior cingulate cortex are differentially activated (Schilbach *et al.* 2013B, 402). Thirdly, they find distinct patterns of neural activity associated with the different roles agents may have when jointly attending to an object. *Following* someone's gaze directed at an object differentially recruits the medial prefrontal cortex, while *leading* someone's gaze recruits the ventral striatum (Schilbach 2015, Schilbach *et al.* 2010, 2702).

Interpreting these neural data, they suggest that *leading* gaze may have a rewarding effect and may lead to an increase in motivation (Schilbach *et al.* 2010, 2013A).[13] Along the same lines, they hypothesize that 2^{nd} person interaction is marked by heightened emotional engagement (Schilbach *et al*. 2013A, 396).

[12] This method allows to obtain eye tracking data from "participants inside the MR scanner to make a virtual character's gaze behavior responsive to the participant's gaze in real time" (Schilbach 2014).
[13] Redcay *et al*. (2013, 435) lends further support to the idea that self-directed gaze in interactions and self-directed gaze from a video replay is associated with distinct neural activity.

Furthermore, using a stimulus-response compatibility task,[14] Schilbach *et al*. (2011) show that gaze shift of an interacting social stimulus influences action control in normal functioning subjects, but not in subjects with high-functioning autism. This indicates that action control in normal functioning subjects is dependent on interactive gaze.

Notably, on Schilbach's account, "social cognitive" differences in interactive and observational situations are wide-ranging as they bear on attention allocation, reward experience, motivation, and action control. It is for these reasons that they speak of a "second person mode" of social cognition that is "fundamentally different" (Schilbach 2014) from third person social cognition. Nevertheless, the differences between these "modes" of social cognition should be understood as a matter of degree.

A second prominent idea concerns the role interaction may play in learning processes. Famously, György Gergely and Gergely Csibra argue for a human-specific learning mechanism which is sensitive to interaction-specific ostensive signals (e.g. eye contact, eyebrow raising) (Gergely 2010; Csibra and Gergely 2006). In these ostensive contexts, according to Csibra and Gergely, the learner is biased to interpret communicative gestures as transmitting generic knowledge about referential kinds (Csibra and Gergely 2009) (rather than just episodic facts). For instance, in one crucial study (Yoon *et al*. 2008) preverbal infants

[14] In this experiment, neurotypical subjects and subjects with high-functioning autism had to produce spatially congruent and incongruent motor responses in response to either a gaze shift of social stimuli or a shift of an object stimulus.

encode information about an object's *identity* in a communicative context (involving eye contact, and infant directed speech); and they encode information about an object's *location* when such a communicative context is absent.[15]

However, Csibra and Gergely's proposal remains controversial when interpreted as a claim specifically about interaction. On the one hand, although children's sensitivity to different types of information may depend on a context being communicative, it remains to be shown whether it also depends on communication *in interaction*. In Yoon *et al*.'s experiment (Yoon *et al*. 2008) a communicative, interactive context was contrasted with a non-communicative, non-interactive context. Hence, it was not established how infants would have responded had they merely observed a communicative context. Secondly, the role of interaction in learning might be more complex than Csibra and Gergely's model predicts. For instance, Shimpi *et al*. (2013) find that imitative learning of novel actions is sensitive to toddler-directed ostensive cues *only if the interactor is familiar to the infant*.[16]

A third proposal concerning social cognition in interaction comes from Henrike Moll and Michael Tomasello (Moll *et al*. 2011a) who argue that children often overestimate the amount of knowledge that is shared between her and the person interacted with (Moll *et al*. 2011a, 256). In their study, two-year-olds first played

[15] In this looking time experiment, infants are shown to be more surprised when an object unexpectedly changes its identity after an actor had pointed to the object in a communicative context. Furthermore, infants are shown to be more surprised when an object unexpectedly changes its location after an agent had grasped the object in a non-communicative context.
[16] In this experiment, 18 month old infants are presented with novel actions (e.g. ringing a doorbell using one's forehead) after a brief warm-up period involving a sorting game. Shimpi found that imitative learning crucially depends on whether the person interacted with later on was familiar from the warm-up period.

with an adult using two toys. Subsequently, in the "Silent Absence Condition", the adult left the room and stopped the interaction with the child. Then a third toy was introduced to the child in the adult's absence. Then the adult returned. Alternatively, in the "Communicative Absence Condition", the adult left the room but kept communicating with the child in her absence from behind a shelf saying things such as "Oh, how nice! Great! Super!" (see Moll *et al*. 2011a, 256). Moll *et al*. found that in the Silent Absence Condition all infants knew that the adult had not encountered the third toy which was introduced in the adult's absence. In contrast, in the Communicative Absence Condition children found it significantly harder to tell which object was unknown to the adult. In light of this finding, Moll *et al*. hypothesize that, in interactive contexts, young children assume that they share the space around them. However, whether this finding points to a cognitive feature specific to interactive engagement has yet to be empirically determined by introducing more controls in the study. For instance, Moll *et al*. did not rule out whether children would rely on the 'shared space' assumption when merely observing an interactive situation.

Notably, these three proposals don't explicitly address the role of belief-desire attribution in interaction and observation. A fourth proposal, the one I will focus on in this chapter, has been coined "enactivism". Enactivists specifically deny that most interactions involve the attribution of beliefs and desires to other agents (an ability I will call "mindreading"). More specifically, a number of enactivists have argued that mindreading should be relegated to the 3^{rd} personal (i.e. observational) contexts (e.g. Hutto 2004; Gallagher 2001; Reddy 2008). The 2^{nd} personal stance

(i.e. the interactive stance) in contrast is, in some important sense, devoid of mindreading. This is not to be understood as a developmental claim alone. Allegedly, interactions between adults don't involve mindreading either. Consequently, enactivists have prided themselves on offering an alternative to the more traditional belief/desire-based approaches to social cognition, i.e. simulation theory (ST)[17] and theory theory (TT)[18] (In what follows, I will refer to theorists who defend one of those theories collectively as "ToMers"). If the enactivist's assessment should turn out to be correct, then ToMers are thoroughly undermined, because, of course, they aspire to accurately capture the cognitive mechanisms underlying real life interactions (see Carruthers 2009, 167).

In short, among other things, enactivists have defended the following two ideas:

Core Thesis. Successful interactions are neither driven nor explained by the interactors' ability to mindread.

Distinctness. The mechanisms enabling 2^{nd} personal social cognition and those enabling 3^{rd} personal social cognition are distinct.

In Section 2, I will clarify what is meant by 'interaction'. Next (Sections 3 and 4), I will examine two enactivist defenses of **Core Thesis**. The first defense was developed by Hanne De Jaegher and Ezequiel Di Paolo who argue that interactions

[17] According to simulation theory, attribution of mental states to other agents is achieved using one's own mental states to simulate the mental states of other agents.
[18] According to theory theory, the attribution of mental states to other agents is achieved through the application of a 'theory'.

cannot be explained in terms of mindreading, because, quite generally, interactions cannot be explained in terms of individual agents' contributions to interactions. Next, in Section 4, I will critically discuss Shaun Gallagher's and David Hutto's account of **Core Thesis**. Both argue that lower level cognitive mechanisms not involving mindreading suffice to navigate interactions. In Section 5, I will argue against Gallagher's and Hutto's view concerning **Distinctness.**

Before starting my discussion proper I'd like to motivate my skeptical stance towards enactivism by discussing two general worries. These worries will cast initial doubt on the idea that social cognition in interactive and observational contexts is fundamentally distinct. First, in the real world, the boundaries between 2^{nd} and 3^{rd} personal contexts simply aren't clear-cut enough in order to merit such a sharp theoretical distinction. Suppose, for instance, a child, Johnny, is interacting with his brother in order to plot something against their sister, Mary, who plays with her toys in the other end of the room. Suppose Johnny now turns to Mary and starts interacting with her. The enactivist might say that Johnny first had a purely observational attitude towards his sister; thereafter he adopted an interactive attitude towards her. But how plausible is this? In the real world, interactions with and observations of others are tightly interwoven; so tightly indeed that it would seem surprising if distinct *theories* were to apply to both contexts. Johnny might observe his sister for a few seconds, then interact with her for a minute, then turn back to his brother observing her again. In the face of such a tightly knit juxtaposition of interactive and observational contexts we might generally tame our expectations concerning hard and fast distinctions between both contexts.

Second, many inferences that can be drawn specifically *in* interactions can likewise be drawn when *observing others interact*. Let me give an instructive example from Stephen Butterfill (see 2013a): He argues that in interactive situations involving joint actions, it is especially easy to correctly attribute intentions to the other agent; this is true simply because joint action often requires sharing of intentions. Suppose I intend to put a stroller on a bus and you help me carry out my plan. If our joint action is to be successful you should, by default, also intend to put the stroller on the bus. If you, say, intend to flip it over, or take its wheels off we won't succeed in carrying out my plan. Butterfill writes that in these situations interactors "may be in a position to know that the goals of her target's actions will be the goals of her own actions" (Butterfill 2013a, 22). Hence, interactions of this form make effortless attributions of goals to others possible. However, at the same time Butterfill recognizes that observers could also acquire such knowledge. While interactors can rely on the "my-goal-is-her-goal" inference, observers could respectively rely on the "her-goal-is-his-goal" inference (Butterfill 2013a, 20). Just as I am in a position to know your intentions when we are carrying out a joint action, so can an onlooker know your intentions when she sees us carrying out a joint action (given she knows my intentions). Interactors do not enjoy a principled privilege. Therefore, at least in this case, alleged genuinely "interactive" features of social interactions can, in principal, be exploited from a third person perspective. Of course, if we put a stroller on a bus I will usually be more motivated to know what your intentions are. Furthermore, I will pay closer attention to what you are doing than as if I were just watching the scene unfold. Therefore, the fact that there

is no principled advantage interactors enjoy leaves untouched Schilbach's claims about the role of motivation, attention, and reward feelings in interactions (see above).[19]

2. Interaction – paradigms without a definition

ToMers and enactivists disagree about whether normal subjects (need to)[20] rely on mindreading in order to successfully navigate social interaction. But what are interactions? In her article "Embodied cognition and mindreading" (Spaulding 2010) Shannon Spaulding gives us a concise example of a prototypical interaction:

> Suppose Jack and Jill are sitting in a coffee shop; both are doing some work on their respective computers when suddenly Jack starts asking Jill questions such as "What are you working on?", "Where are you from?" etc.. When Jill gives only cursory answers such as "philosophy", Jack responds "Oh, I bet you are really deep" to which Jill just responds "sure". This goes on for a little while but, ultimately, when he realizes that Jill won't reciprocate the way he'd like, Jack lets Jill off the hook and they both go about their work.

If ToMers and enactivists disagree about what enables agents to successfully interact, they surely disagree about the kind of situation just described. ToMers will most likely analyze this situation along the following lines: Jill believes that Jack

[19] Furthermore, research concerning social cognition in high-functioning autism indicates that there is *some* difference between social cognition in interactive and observational contexts. Schilbach *et al*. (2013) hypothesize that it is specifically social cognition in interaction which may be impaired in high-functioning autism.

[20] Recent studies concerning the automaticity of mindreading (e.g. Qureshi *et al*. 2010; Schneider *et al*. 2014) seem to indicate that others' mental states may be computed and attributed to others even in situations in which this is not at all necessary. Therefore, the claim that mindreading is necessary for social interaction needs to be distinguished from the claim that mindreading is, in fact, employed.

believes that she is interested in him. She also believes that Jack's belief is false, because, in fact, she is not interested in him. Jack initially believes (or hopes) that Jill has a desire to talk to him, but when Jill keeps giving curt answers he finally realizes that this belief was false (see Spaulding 2010). Hence, ToMers believe that what drives social interaction is mindreading. Enactivists reject this interpretation.

Definitions of the relevant terms '2nd personal stance', '3rd personal stance', 'interaction', and 'observation' etc. are hard to find in the literature.[21] However, I believe that, ultimately, it is not necessary to provide such definitions, which would be called for if there were vast disagreement about which situations are interactive, observational etc.. But this is not the case. Philosophers and cognitive scientists by and large agree about which cases they are disagreeing about. The disagreement is about the *correct analysis* of the relevant cases, not about their identity.[22]

It is worth noting that the interaction/observation distinction cannot be sufficiently identified using a grammatical criterion. The different stances don't map onto the grammatical distinction between the use of the personal pronouns "you" (for the interactive stance) and "he", "she", "it", and "they" (for the observational stance). Simply put, there are interactions in which we don't use any

[21] Notably, De Jaegher and Di Paolo (2007) is an exception. They give the following definition of "interaction": Social interaction is the regulated coupling between at least two autonomous agents, where the regulation is aimed at aspects of the coupling itself so that it constitutes an emergent autonomous organization in the domain of relational dynamics, without destroying in the process the autonomy of the agents involved (though the latter's scope can be augmented or reduced) (De Jaegher and Di Paolo 2007, 493).
Following an interpretation by Herschbach (2012), "coupling" amounts to the coordinated mutual dependence of the behavior of several subjects. Furthermore, coupling can be said to be "*regulated*" if "engaging in motivated changes to the constraints or parameters that influence the coupling" (Herschbach 2012). One worry concerning De Jaegher's definition is that it might not accurately distinguish interaction from mere coordination. Arguably, coordinated action also requires regulated coupling between autonomous agents.

[22] Note that other related discussions may well benefit from definitions of these terms. Categorization of the relevant cases is far less obvious when, say, comparing interaction to cooperation. After all, it is not intuitively clear which cases exemplify cooperation and which ones exemplify interaction. It is just that in the present discussion these definitions are not necessary.

of these pronouns. Second, the use of these pronouns doesn't tell us anything about the nature of the respective stances. They could, if anything, just provide a marker.

3. The Constitutivist account of "Core Thesis"

In a series of articles Hanne De Jaegher has developed the view that interaction is constitutive of social cognition[23] (e.g. De Jaegher *et al*. 2010; De Jaegher and Di Paolo 2007); i.e. interaction is an essential proper part of social cognition. This account is intended to stand in opposition to traditional ToMistic views according to which social cognition is reducible "to the workings of individual cognitive mechanisms" (De Jaegher *et al*. 2010). De Jaegher argues that "interactive processes [...] complement and even replace individual mechanisms" (De Jaegher *et al*. 2010). Her view can be summarized as follows:

> **Constitutivism about Interaction.** Central features of social cognition in interactions cannot be explained solely in terms of each interactor's contributions to these interactions (e.g. their behavior and their mental states). Rather, interaction is explanatorily basic for social cognition.

Note that De Jaegher's claim is quite radical in that she does not merely hold that features of interaction can *causally* influence an individual's cognitive processing (which would be uncontroversial).

A general problem with her approach, pointed out by Herschbach (2012), is that

[23] "[S]ocial cognition", in this context, is defined as a "[g]eneral term used to describe different forms of cognition, about, or actions in regard to, agents or groups of agents, their intentions, emotions actions and so on, particularly in terms of their relation other agents and the self" (De Jaegher *et al*. 2010).

her constitution claim involves a category mistake. Herschbach asks "[i]n what sense could a social interaction be a constitutive element of a *neural mechanism*?". And he continues "[i]f a constitutive element is understood as a 'part of the phenomenon' itself, this statement would involve a substantial confusion between levels of organization" (Herschbach 2012, 477). The worry, put more colloquially, is that interaction is something that goes on *between subjects* while neural processes are something that happen *inside one subject*. One may worry that this critical assessment overextends in that it would amount to a *general* criticism of the extended mind thesis (i.e. the claim that cognition is not confined to what's going on beneath the skull). Hence, rejecting De Jaegher's claim would require a more thorough treatment of the extended mind literature.

A more immediate problem for De Jaegher's constitution claim is that her examples don't unambiguously support her case. For instance, she relies on a "perceptual crossing" experiment (De Jaegher *et al*. 2010) conducted by Auvray *et al*. (2009). In this experiment, two blindfolded participants interact with each other by moving an avatar along a one-dimensional strip. For each player's avatar there is also a shadow that replicates the avatar's movements. Furthermore, along the strip there is an additional static object. Hence, each player can encounter three different objects: The other player's avatar, the other player's shadow, and the static object. When a player encounters any object she receives sensory feedback.

For all three objects, the sensory feedback is identical; hence, a player cannot distinguish between the different objects encountered by relying on sensory

information alone. Crucially, when two avatars meet, *both* players receive sensory feedback. When an avatar encounters a shadow only the player whose avatar it is receives feedback.

De Jaegher notes that "in such an impoverished sensory situation, participants find each other and concentrate their mouse clicks on each other's sensors (65.9 % of clicks) and not on the identically moving, but non-contingent shadow objects (23 %)" (De Jaegher *et al*. 2010, 444). The players' "finding each other" is explained by their behavior when they encounter an object. Upon receiving sensory feedback, a player tends to reverse her direction. When an avatar encounters a shadow, the avatar reverses direction, but the shadow does not. When two avatars meet, they both tend to reverse direction and start oscillating around each other. According to De Jaegher, this experiment provides evidence that the agents' finding each other can't be explained by each player's contribution. This is because each player is inept to even distinguish a shadow from an avatar.

De Jaegher's interpretation can be resisted. The perceptual crossing experiment does not establish that interaction is basic or *constitutive* of social cognition. Granted the experiment *does* show that not all features of the interaction can be explained by just looking at *one* player's contribution. However, everything that happens in Auvray's experiment is entirely predictable if we take into account *both* players' contributions. The interaction effect De Jaegher describes is fully determined by the pattern of sensory feedback that each player receives in conjunction with their individual strategies (i.e. reversing the avatar's direction upon

receiving sensory feedback). The idea, however, that all facts about interactions cannot always be explained solely in terms of *one* interactor's contribution is not very controversial. Suppose, for instance, you and I want to put a stroller on a bus. What explains our success in completing this task? Surely, a satisfactory explanation would need to appeal to both of our goals; it would need to take into account that your actions (say, the speed with which you lift the stroller) have an effect on what I do (say, lifting the stroller with equal speed). ToMers can embrace the idea that each interactor's actions have a *causal* effect on the respective other's cognition. However, what these examples do not show is that interaction is *constitutive* for social cognition. I conclude that the relevant data in support of De Jaegher's view *can* be explained within an individualistic paradigm.

In this section, I argued against De Jaegher's view that interaction cannot be understood in terms of individual agents' cognitive mechanisms. My main line of reasoning was that the constitutivist approach lacks convincing examples. In the next section, I will focus on the more moderate enactivist theory developed by Shaun Gallagher and Daniel Hutto who argue that, although interaction is not "basic" in the aforementioned sense, successful interaction does not require mindreading.

4. Shaun Gallagher's and Daniel Hutto's account of "Core Thesis"

Shaun Gallagher and Daniel Hutto defend a type of two-systems account of social cognition. System 1 operates fast and unconscious. It does not involve mindreading, but, rather, exhaustively recruits lower-level mechanisms which

Gallagher and Hutto label "primary" and "secondary intersubjectivity" (PIS and SIS). PIS enables agents to interact with *one another*; SIS enables agents to adopt a shared perspective with regard to *the world*. PIS comprises cognitive mechanisms such as "gaze following", "emotion detection", and understanding of goal-directed actions. SIS comprises cognitive mechanisms such as "joint attention" and the ability to understand others' emotionally valenced attitudes towards an object or a situation (see Gallagher 2001, 2008a, 2012; Hutto 2004).

System 2—the mindreading system—is slow, non-modular and solely consciously employed. Think, for instance, of the reasoning underlying Sherlock Holmes's painstaking reconstruction of the murderer's motive. Crucially, such conscious reasoning about mental states is not fast enough to guide and direct interaction. Typically, Gallagher and Hutto maintain, in interactions there is no time for a slow and cognitively costly reconstruction of the other agent's mental states. Therefore, agents have to rely on low level cognitive mechanisms PIS & SIS.

Furthermore, mindreading is supposed to be constitutively 3^{rd} personal (Hutto 2004). Hutto explains that "[w]e ascribe causally efficacious inner mental states to them [other agents] for the purpose of prediction, explanation, and control" (Hutto 2004, 549). This amounts to viewing them as "foreign bodies" (ibid., 549) and, thereby, taking a spectatorial stance towards them. In contrast, when we interact with other agents we rely on more basic forms of "primary" and "secondary intersubjectivity" (see Gallagher 2012, 2001, 89; Hutto 2004, 550).

Paradigmatically, we take a 3^{rd} personal stance towards other people when their

actions seem unfamiliar and atypical to us (see Gallagher 2001, 92; Hutto 2004). In such cases, we theorize about others' beliefs and desires in trying to explain their actions. However, when everything goes as usual mentalizing is unnecessary.[24]

System 2 reasoning about mental states is essentially conscious. Therefore, few will deny its existence, and I won't spend time on this part of the enactivist theory. Disagreement arises with regard to the enactivist's system 1. Is it true that the unconscious cognitive processes which guide and regulate swift social interaction do not employ mindreading? Do system 1 mechanisms SIS and PIS provide sufficient cognitive resources to explain successful social interaction?

Let me discuss what I take to be the two most pertinent and contentious issues. First, infants reliably pass interactive, non-verbal false-belief tasks at 18 months of age and younger. It is clear that at this young age, infants couldn't possibly *consciously* reason about false beliefs, or rely on narratives to guide their understanding. Hence, the enactivist is called upon to give an ersatz-explanation that does not rely on mental state attribution. Second, Hutto and Gallagher adduce a principled argument for why mentalizing *cannot* drive social interactions. This argument states that, by ToMers own lights, mentalizing is used to "predict and explain" others' behavior. However, predicting and explaining behavior couldn't possibly be an unconscious process. If mindreading is a conscious process, then it could not underlie interaction (for reasons stated above). I will discuss both issues

[24] If, in a given situation, these low level cognitive tools don't suffice, according to Hutto, narratives help us become familiar with social situations (e.g. Hutto 2009).

in turn.

Firstly, in a study by Buttelmann *et al*. (2009), 18 months old infants succeed in helping an adult retrieve an object from a box, while, according to the standard interpretation, taking into account the adult's false belief about the object's location. In the experiment, an infant watches how an adult sees an object being placed into one of two boxes (box A). Then, in the false-belief condition, the object is moved from box A to a different box (box B) in the adult's absence. When the adult finally tries to retrieve the object from box A (due to her false belief) the child helps the adult, leading her to box B which contains the object. The infant, however, only helps the adult retrieve the object in the false-belief condition. In the true-belief condition in which the adult knows the true location of the object and yet still opens the empty box, the child helps the adult open the empty box assuming that she must have some other reason to open it.

According to the mentalizing interpretation of the active-helping study (which Gallagher rejects), the infant understands that, in the false-belief condition, the adult *believes* that the object is in box A, and that she *wants* this object. This is what motivates the infant to help. According to a different, non-mentalizing interpretation (usually labeled "the behavior rule interpretation"), the infant knows a rule such as "people look for objects where they last saw them". Rules such as this one are meant to enable the infant to, say, distinguish situations in which the adult looks for an object from situations in which she does not look for it. This, in turn, is important for knowing when to help and when not to help. Gallagher rejects both of

these interpretations, arguing that there is a distinctively enactive way of viewing these findings. He states:

> [...] the fact that the infant knows either that the agent has been in a position to see the switch or not, plus the agent's behavior with respect to A [...], is enough to specify the difference in the agent's intention. For the infant, that signals a difference in affordance, i.e., a difference in how the infant can act, and thereby interact with the agent. The infant does not have to make inferences to mental states since all of the information needed to understand the other and to interact is already available in what the infant has seen of the situation (Gallagher 2012, 201).

And

> The phenomenological-enactive approach provides an alternative to both the ToM and behavioral interpretations (Gallagher 2012, 202).

Notably, in this passage, Gallagher focuses on knowing "intention[s]" (and not on knowing behavior). Accordingly, the enactivist may depart from a behavior rules account of social cognition by relying on intention rules: "people *intend* to look for objects where they last saw them". This interpretation, very much in the behavior-rule spirit, introduces a further complication: Once the infant knows the adult's intentions, she then has to employ the additional rule "people tend to do what they intend". Therefore, by putting intentions in the focus of analysis, the enactivist cannot hope to get around a behavior rule which maps intentions to actions. Furthermore, it is not clear what the motivation for such an 'intention-rule' could be.

One of the attractions of behavior rules is their alleged parsimony (they don't involve mentalizing of any sort). Intention-rules, on the other hand, *do* involve mentalizing (they involve *intention* attribution); hence such rules would be *less* parsimonious, and, therefore, we'd be owed an account concerning the benefits of such rules.

A second, distinctively enactive perspective concerns the explanatory role of agents' possibilities for action. Standardly, ToMers hold that the infant's action possibilities (e.g. the possibility to help the adult open the box) are *grounded in* an understanding of the situation; in an understanding that the adult wants the object and that she has a false belief about its location. Alternatively, avowed enactivists sometimes hold that the direction of explanation should be reversed: Action possibilities sometimes ground how objects and situations are represented. This line of reasoning is famously adopted by Alva Noë who argues that certain properties of perceptual content are constituted by sensory-motor relations. His most thoroughly discussed example concerns the perception of a tomato (see Noë 2008). He starts with the following observation: Looking at a tomato we can only literally see one side of it (the side facing us). Nevertheless, we *perceive* tomatoes as three dimensional objects. We perceive it as an object which has some hidden sides. According to Noë, the perceptual content of a tomato as a three dimensional object is *constituted by* or *grounded in* the "availab[ility] to perception through appropriate movement" (Noë 2008, 16). The tomato's hidden sides are present in perception, *because* upon perceiving one side one has motor access to a visual representation of its hidden sides.

This line of reasoning is reflected in the following quote from Gallagher: "[I]nfants understand others *in terms of how they can interact with them*" (Italics by the author) (Gallagher 2012). Hence, understanding others' mental states may be grounded in possible ways to act and interact with the agent. However, without taking a stance on Noë's view on perception, this line of reasoning is hardly plausible for the relevant social cognition cases. To see this, reconsider the active helping study. Suppose that the infant's understanding of the adult's beliefs and desires were based on her grasping that it is appropriate to help in one situation but not in the other. What, then, explains the child's sensitivity to situations in which helping is (or is not) appropriate? Surely, it cannot be the attribution of beliefs and desires. But grasping action possibilities cannot be *bare* either, simply because different situations afford different actions; and first agents need to understand a situation in order to know which actions are afforded.

Noë's perception-based examples and the "social-cognitive" paradigms are disanalogous in an important sense. On the one hand, we have lots of experience perceiving, handling, and modifying objects. This is what supposedly grounds sensory-motor expectations. We know that objects such as tomatoes will reveal hidden sides when we go around them and when we move them in our hands, because we have seen this happen many times before. It is not clear what the relevant prior experience in the active helping study would be. Surely, we have *vastly* more experience discovering hidden sides of three-dimensional objects than we have with others' false beliefs; especially at 18 months of age. The situation in which an object is moved from one box to another while the adult is absent is

comparatively unique for the child.

Yet another line of argument frequently adopted by enactivists relates social cognition to "direct perception" (Gallagher 2008, 536). According to Gallagher, we "have a direct perceptual grasp of the other person's intentions, feelings, etc." (Gallagher 2008). Intention attributions are therefore not *mediated* by either a theory or a behavior rule, which means that "there is no problem of other minds" (Gallagher 2008). Because mental states can be perceived directly, there is simply no need for any intermediate cognitive mechanisms.

The 'immediate perception' view is problematic if intended to provide an *alternative* to mentalizing accounts of social cognition. This is because questions concerning the contents of perception and questions concerning the underlying mechanisms of social cognition should be kept distinct. The basic argument is this: If perceptual content is conceptual at least in some cases, then it is a live option for theory theorists to argue that the conceptual outputs of a mindreading module can be a constitutive part of perceptual states (see Carruthers 2015). Of course, this would not provide an immediate answer to the question whether *mental state concepts* can figure in perception. However, it would guarantee the in-principle compatibility of theory and a direct perception account of mental states.

Let me give two arguments in favor of the view that concepts can be part of perceptual states. The first argument is phenomenological in nature, the second draws on the tight interplay between perception and cognition. In a recent article, Carruthers (see Carruthers 2015) suggests that when we see something, S, as an

instance of a kind, K, the concept that represents K is "bound into" the perception of S (Carruthers 2015, 6). For an illustration, think of perceiving a cloud. You stare at a cloud when all of the sudden you realize that it looks face-like (or, say, wardrobe-like); i.e. you see the cloud *as* a face, or *as* a wardrobe. In such cases, Carruthers argues, the concept FACE is bound into the perception of the cloud. Along the same lines, theory theorists could argue that the conceptual outputs of a mindreading theory module can be "bound into the contents of the perceptual states that provide the basis for its interpretations" (Carruthers 2015, 7). Such examples are persuasive, because it is hardly plausible that the mental representations FACE, or WARDROBE are essentially non-conceptual.

Secondly, a growing body of literature indicates that there is a tight interplay between conceptual knowledge and visual perception. One plausible explanation of this interplay is that perception just is conceptual. Evidence comes mainly from research concerning links between color perception and color concepts (e.g. Thierry *et al*. 2009; Winawer *et al*. 2007; Daoutis *et al*. 2006). One experiment by Daoutis *et al*.'s (2006) involved 4–7 year-old children from either England or Kwanyama (Namibia). The crucial difference between both groups was that the Kwanyama don't have distinct color terms for the colors blue and green, blue and purple, and red and pink. In the experiment, the children had to find a target color in an array of color patches which contained patches of either the target color or distractor colors. The distractor colors were designed to be either cross-category for English speakers, and within-category for Kwanyama speakers; or, in a second condition, cross category for both groups. Daoutis *et al*. found that within-category search was

faster for the English speakers. This effect did not hold for the Kwanyama speakers (for whom there was no within-category condition). One attractive (but not the only available) interpretation of the data is that color concepts form a part of color perception. This would explain why differences in conceptual knowledge predict performance in visual search tasks.

Carruthers considers an alternative explanation of these findings according to which "concept acquisition permanently "warps" the processing that takes place in midlevel visual areas" (Carruthers 2015, 9). However, he argues that long-term "warping" is unlikely, because interaction effects between color perception and color concepts are highly sensitive to online interference effects. Typically, the concept-based performance differences in these tasks go away under cognitive load. These arguments don't conclusively settle whether or not perception itself is conceptual (or whether conceptual knowledge merely has causal effects on perception). But my goal is more moderate. I showed that ToMistic accounts of social cognition are in principle able to embrace a direct perception account of social cognition. It is at the very least a live option for ToMers to hold that perceived mental states could be the result of a complex interplay between conceptual mindreading systems on the one hand and perceptual systems on the other.

Let me now go on to discuss enactivist claims concerning the role of prediction and explanation of behavior. Enactivists have argued that folk-psychological attributions of beliefs and desires serve to *causally predict and explain* behavior

(Gallagher 2001, 102; Gallagher 2012; Hutto 2004, 549). According to enactivists, ToMers share this view.

The claim that what one is doing when mindreading is *explaining* or *predicting* the other person's action in terms of mental states, however, is not *my* claim. It's a claim that is pervasive in the ToM literature (Gallagher 2012, 205).

Similarly, Hutto writes "it is also generally assumed [by ToMers] that we are normally at theoretical remove from others such that we are always ascribing causally efficacious mental states to them for the purpose of prediction, explanation and control" (Hutto 2004, 548).

Allegedly, this particular mode of understanding others leads to "estrangement", and, therefore, it cannot serve as the right model for understanding others in interactive contexts. Hutto argues that predicting and explaining others' actions is only necessary when actions are unfamiliar to us. In most circumstances, however, "we already know what to expect from others and they know what to expect from us in familiar social circumstances" (Hutto 2004, 558).[25]

Now, "know[ing] what to expect" does not absolve us from *predicting* what others do. Suppose, for instance, you bump into somebody in the hallway whom you want to pass. Suppose a convention exists according to which, in these

[25] According to Hutto, one reason to favor narrative-based accounts over ToMistic accounts is its phenomenological accuracy. We simply don't go around consciously calculating others' beliefs and desires all the time. However, it is doubtful that narrative-based models fare better with regard their phenomenological accuracy. As it is, we also don't go around recalling stories that might fit a particular interactive situation. Understanding others is often *entirely* effortless. Therefore, any theory about social cognition which gives lots of weight to phenomenological 1st person data will have to refrain from positing *any* explanatory mechanism. This, however, seems implausible. As John Michael argues, surely, in understanding others, interpretation has to happen somewhere (see Michael 2011, 562).

situations, both people step to their respective right. Surely, in this case, you know what to expect from the other person: she will take a step to her right. You take a step to your right and, hence, you both succeed in passing each other. You both knew what to expect, because you both knew the pertinent rule for such situations. However, the fact that this coordination problem was particularly effortlessly and easily solvable does not mean that you didn't have to predict what the other person would do. You predicted that she would take a step to the right and that is why you stepped to the right. Of course, this does not entail that the enactivist's analysis is wrong. In fact, there is little reason to assume that solving the hallway problem involves mindreading. All it shows is that enactivists will also have to appeal to the prediction of others' behavior at some level in their theory. In the hallway case, the prediction might have been facilitated by the existence of a social convention to step to the right, which may have obviated the need to mindread. Social predictions and explanations can be accomplished in various ways (e.g. through mental state attributions, social conventions, or behavior guiding rules). They don't always involve mindreading. However, the view that ordinary social circumstances don't require any predictions is flawed and ToMers' positions cannot be ruled out on those grounds alone.

A similar argument can be given for 'explanation'. Actions are often ambiguous; one and the same physical action can mean different things and can be interpreted in various ways. Reconsider, for instance, the active helping study. When the agent comes back and tries to open a box, there are several things she could be interpreted as doing. She may be trying to open the box, lift the box, or break off the handle.

Surely, an adequate understanding requires ruling out some of these possibilities. Ascribing causally efficacious beliefs and desires is one way to reach adequate understanding. According to this model, the child understands what the agent does when she knows that the adult wants the object and that she has a false belief about its location.

Enactivists point out that disambiguation does not always require mentalizing. Rather, understanding is achieved by certain behavioral scripts and lower level cognitive mechanisms. For instance, suppose you stand at the register in the super market. The person behind the register reaches towards you. The display reads $10.53. The appropriate action in this context is to hand her the money. How did you know that this would be the appropriate thing to do? One possibility is that you understood that she *wanted* money from you (and you owe the money). An alternative explanation is that acting in this way was just demanded by the situational setting. Whichever description turns out to be right, it is clear that there needs to be some disambiguating explanation of why the person behind the register acted the way she did. Hence, the ToMistic view cannot be ruled out on the grounds that they provide some such explanation.

In this section, I argued that the enactivist defense of **Core Thesis** does not provide a genuine alternative to more conventional accounts of social cognition. On my interpretation, Gallagher and Hutto's views are close to a behavior rule account of social cognition which is then combined with a direction-perception account of mental states. Furthermore, I argued that mentalizing accounts of social cognition

cannot be ruled out merely on the grounds that they involve explanations and predictions of behavior.

5. Gallagher's and Hutto's account of "Distinctness"

Suppose the enactivist were right in that paradigmatic interactions are free of mindreading. In this case, as I will now go on to show, she is forced to give an enactive analysis of some entirely 3^{rd} personal false belief paradigms (hence, undermining the theoretical distinctness between 2^{nd} and 3^{rd} personal paradigms).

Evidence comes from so-called "spontaneous response" tasks (e.g. Onishi and Baillargeon 2005; Surian *et al*. 2007; Woodward *et al*. 2009). In these tasks, children's understanding of others' false beliefs is inferred from "behaviors they spontaneously produce as they observe a scene unfold" (Baillargeon *et al*. 2010). There are two types of spontaneous tasks. On the one hand there are violation-of-expectation paradigms which exploit the fact that an infant will look longer at an agent or a scene, if her actions don't match the infant's expectations. On the other hand, there are anticipatory looking tasks which exploit the fact that infants will look in the direction of a location in which they anticipate others to act. Anticipatory looking can be sensitive to (false) belief attribution, because the infant predicts the agents' actions based on belief attribution. Importantly, spontaneous response tasks are observational paradigms in which the infant merely watches a certain scene unfold.

For instance, in a violation-of-expectation paradigm, Onishi and Baillargeon

(2005) found that 15-month-olds have an understanding of false beliefs. In this experiment, infants were first familiarized with a toy that stands between a green and a yellow box and which is then hidden in the green box. Next, the agent reached inside the green box to retrieve the toy. Next followed a belief induction faze. In the false belief condition the toy was moved from the green to the yellow box while the adult was absent. When the adult reached for the box where she didn't believe the toy to be, infants looked reliably longer than when the adult reached for the ball in a location incongruent with her false belief about the ball's location; hence, taking into account the adult's false belief, the infant expected her to look for the toy where she falsely believed it to be.

It is clear that enactivists cannot readily embrace the false-belief-tracking interpretation. They believe that belief-desire attributions are the product of *conscious* reasoning. Violation-of-expectation paradigms conducted with 15-month-old infants are not the purview of enactive system 2 mindreading. Furthermore, it would seem quite ad hoc to suppose that Baillargeon's violation-of-expectation paradigm on the one hand and Buttelmann's active helping paradigm are in some basic theoretical way distinct. The only motivation for this view would be the defense of **Distinctness.** But, as I said, this seems ad hoc. Hence, the enactivist needs to explain the violation-of-expectation findings relying on enactivist tools (e.g. in terms of primary and secondary intersubjectivity). Though there is nothing interactive about this paradigm; it is entirely observational.

Therefore, if the enactivist is right in that belief/desire attributions are the result

of effortful conscious thought processes, she has to admit that, at least in some cases, enactive social understanding is 3^{rd} personal. This is because infants couldn't possibly consciously reasons about others' beliefs and desires. If, however, the enactivist admits that Baillargeon's paradigm does involve mindreading, then she also has to be comfortable with the idea that mindreading is a largely effortless, unconscious process.

6. Conclusion

Although it is plausible that social cognition evolved in order to navigate social interactions (Carruthers 2009, 167), hard and fast cognitive differences between interactive (i.e. 2^{nd} personal), and observational (i.e. 3^{rd} personal) situations prove not be supported by the evidence. I have discussed four claims in support of this claim: First, in real world scenarios interactive and observational paradigms are tightly interwoven. Second, certain allegedly interaction-specific inferences have 3^{rd} personal counterparts and can therefore also be drawn from an observational perspective. Third, De Jaegher's claim that interaction *constitutes* social cognition is untenable. Fourth, the enactivist idea to relegate mindreading to 3^{rd} personal contexts is implausible.

All told, distinguishing 2^{nd} and 3^{rd} personal contexts based whether they involve mindreading, understood as the attribution of beliefs and desires to other agents, is not plausible. However, social cognition in both contexts may still be distinct in less extreme ways. The growing body of research on the automaticity and spontaneity of

mindreading (see Qureshi *et al*. 2010; Surtees and Apperly 2012, Schneider *et al*. 2014) may shed further light on subtle issues concerning the exact conditions under which mindreading is employed. Moreover, thorough research by Schilbach and colleagues show that interactions involve distinct patterns of neural activation which is associated with motivational, attentional, and reward related "social cognitive" differences. The evaluation of this research would be a further step towards fully understanding whether there is something special about interaction in social cognition.

Chapter 3: Lucky Joint Action

1. Introduction

Traditional accounts of joint action (e.g. Bratman 2013; Miller 2001; Tuomela 2005) comprise at least the following two necessary conditions for joint action to occur. First, an intention condition according to which joint action requires that each agent intend the same interdependent end; i.e. an end whose satisfaction requires that each party enact their respective part. Next, it is argued that these intentions be common knowledge.

Under the banner of "minimal joint action", several authors have recently challenged the common knowledge requirement. Most notably, Olle Blomberg (2016) has argued that, in certain cases, false beliefs about one's co-participants' intentions are compatible with joint action. Stephen Butterfill (2011) made room for the possibility of joint action in young children by simply not including a common knowledge condition in his definition of joint action. Similarly, Cordula Vesper *et al*. (2010) do not include common knowledge among the "building blocks" of minimal joint action. Lastly, Christopher Kutz (2000) has argued that participants merely need to be open to the disclosure of the relevant attitudes; a requirement which is, on Kutz's estimation, weaker than common knowledge.

Now, although common knowledge of the intentions of each is, according to these authors, not necessary for joint action, such knowledge is, in each case, argued to be compatible with joint action. Common knowledge is always good, yet not always required. In this chapter, I will further explore common knowledge failure in joint

action by pointing to cases in which common knowledge of the relevant intentions would act as an underminer for the joint action.

The principal idea is that joint actions permit a certain degree of luck. In the relevant cases, the participants believe that the intentions of each robustly favor the joint activity; this belief turns out to be false; the intentions of each do not, in fact, robustly favor the joint activity. However, had the actual intentions of each been common knowledge, the joint action would not have occurred in the first place. More specifically, I have two types of cases in mind; I will, first, analyze a set of cases in which common knowledge of some of one's co-participants' *subplans* would undermine joint action. In discussing these cases, I will rely on the following vignette:

> **Lucky Jog.** Sarah and Bob both intend that they go jogging. Sarah believes that Bob would continue the jog even if it rained. This is important for her! Her intention that they go jogging is conditional on her belief that he wouldn't bail if it rained. Her belief about Bob, however, who would bail if it rained, is false. Fortunately, sunny weather prevails and they complete a happy jog. As it happens, they got lucky.

Second, I will point to a set of cases in which common knowledge of one's co-participants' joint ends would undermine joint action. Consider the following vignette an illustration of this:

> **Forking Trip.** You and I are in Baltimore. I intend that we go to NYC. As a means, I also intend that we go to Philadelphia. You intend that we go to

> Ocean City. As a means, you intend that we go to Philadelphia. However, I only care about going to Philadelphia with you insofar as it is a means to going to NYC. If I knew that you intended that we go to Ocean City, I would simply fly to NYC. You only care about going to Philadelphia with me insofar as it is a means to going to Ocean City. If you knew that I intended that we go to NYC, you would simply fly to Ocean City. I, however, (falsely) believe that you intend that we go to NYC; you falsely believe that I intend that we go to Ocean City. Upon arriving in Philadelphia, we discover our mismatched intentions. Nevertheless, we jointly went to Philadelphia.

This chapter is a contribution to specifying minimally necessary conditions for joint action. Let me briefly explain why such a project is worthwhile. Most importantly, joint action is a pervasive feature of human sociality; that is, "joint action" refers to a basic way of acting together that is distinguished from mere parallel action (e.g. Kutz 2000, 5). To see the difference between both ways of acting, contrast, for instance, the case of two strangers walking down Fifth Avenue next to one another each intending not to run into the other, and the case of two friends walking together down Fifth Avenue. Joint action theory sets out to explicate the distinguishing features of such examples. As the walking-together example indicates, joint actions are not confined to long-term projects such as building a bridge together (Miller 2001, 75), or going on a trip to New York City together (M. E. Bratman 2013). A range of short-term activities such as lifting a table together, bouncing a block on a trampoline together (Warneken *et al*. 2006), or jointly building a block tower serve as paradigmatic examples. These

examples suggest that joint action is a ubiquitous and fundamental feature of human sociality. Concomitantly, philosophical interest derives (at least in part) from its pervasiveness. The literature chimes with this assessment. Margaret Gilbert describes joint action as the "social atom" that lies at the very "foundation of human social behavior" (Gilbert 2003, 39). Michael Bratman thinks of joint action as the structure that grounds "social coordination and planning" (M. E. Bratman 1993, 99). Yet others argue that a cognitive adaption for joint action is the central and most basic evolutionary ingredient separating "hypersocial" (Tomasello *et al*., 2005) human sociality from the sociality of great apes.

In section two, I will discuss all relevant pieces of terminology and presuppositions. In section three, I will detail two types of assurance in joint action. In section four, I will show why the joint actions exemplified in **Forking Trip** and **Lucky Jog** require that the beliefs of each about the respective other's intentions be false.

2. Terminology and presuppositions

In the introduction, I used the somewhat technical terms "ends", "intentions", "belief", and "common knowledge". Let me clarify and explicate these notions.

Ends and intentions. Joint actions involve some kind of motivating attitude on the part of each participant. This attitude is sometimes called a "conative" attitude. The exact name and nature of the relevant attitude varies across philosophical accounts. According to Seumas Miller, jointly acting agents are said to share the same "end" (Miller 2001, 57). Raimo Tuomela speaks of "aim intentions" (Tuomela 2005).

Michael Bratman speaks of regular intentions (with the special content that "we" do something (M. E. Bratman 2013, 60). Finally, John Searle famously defends the idea of *sui generis* "we-intentions" (Searle 1990). At times, the specific differences between those usages are identified to be merely terminological. According to Raimo Tuomela's assessment (2005, 353), for instance, his "aim-intentions" and Miller's "ends" really amount to the same thing. Other times, the differences between these attitudes are taken to be more substantive. Bratman, for instance, argues that only intentions, but not goals, can be "agglomerated" (i.e. can be combined) (M. E. Bratman 2013, 22). Such differences won't matter for our purposes. For the duration of this chapter, I will mostly use the term "intentions" and "ends" to refer to the relevant conative attitudes. I will say that several agents "intend that ___" and, when necessary, use the more specific formulation "have as an end that ___."

Contents. I will use Bratman-style formulations to specify the contents of the relevant conative attitude; these contents have the form "that we ψ." Hence, I will write "you and I have as an end that we go to NYC"[26] or "you and I intend that we go to NYC." Using Bratman-style "that we ψ" formulations in part just fixes a convention. However, one not entirely conventional merit is that such formulations succinctly capture the idea that there is interdependence between the participants' roles. For instance, my intention that we go to NYC cannot be satisfied unless you also go. This is because, on Bratman's account, you enacting your role is part of the satisfaction condition of my intention. The idea that joint actions involve some such interdependence is widely shared (e.g. Tuomela 2005, 340; Miller 2001, 56; Bratman

[26] Such formulations can be found in Miller (2001, 75).

2013, 65). Bratman-style formulations capture this idea in a succinct and intuitive way. In what follows, I will call ends/intentions with a "that we" content *interdependent* end/intentions.

Means and Ends. Some intentions are mere means, some are mere ends, and some are both. The relevant distinction between means and ends is that, other things equal, we only care about the means insofar as we care about the end. Hence, giving up the end would, likewise, rationalize giving up the means. The specification "mere" is important, because it is, of course, possible to intend something as a means *and* as an end. To see all this more clearly, reconsider **Forking Trip**. In this example, our intention that we go to Philadelphia depends on our intention to go to NYC/Ocean City. We each only care about going to Philadelphia insofar as we care about going to NYC/Ocean City with the respective other. This dependence marks our intention that we go to Philadelphia as a mere means. This is different from the case in which going to Philadelphia with you has independent appeal; in which case this intention would also be an end.

Famously, Michael Bratman distinguishes between interdependent ends and the means – "sub-plans" (e.g. Bratman 2014, 55) as he calls them – that are realized in order to satisfy these ends. Ends and subplans are not merely distinct insofar as they are structured by the means-end relation. Subplans and ends are also said to be governed by distinct theoretical requirements. Joint action, according to Bratman, permits, for instance, that one leave the relevant sub-plans somewhat unspecified. All that is required for joint action is that one *intend* that the relevant subplans overlap (or

"mesh" in Bratman's terminology) (see Bratman 2013, 53). Such openness, however, is not permitted for the intended ends. Bratman requires that all parties *have* the same interdependent end, not merely that they *intend* to have the same end.

Note that Bratman merely requires that the *interdependent* ends of each agent coincide. He does not require that *individual* ends of each participant coincide. Shared intention, according to Bratman, does not "require that the agents participate in the pursuit of the same goals. Perhaps you participate in our shared intention to paint the house because you do not like the present color, whereas I participate because I want to get rid of the mildew" (Bratman 2013, 29). It makes sense that joint action should not require our individual ends to coincide. To illustrate, suppose you and I intend that we go to NYC together. I, however, plan to go on to travel to Boston; Boston is my final destination. The fact that my ultimate end is Boston and not NYC does not undermine the possibility for joint action.

These distinctions are important, because, as announced above, one claim of this chapter is that, in some cases of joint action (exemplified by **Forking Trip**), it is required that agents misrepresent their co-participant's *interdependent* ends; not their individual ends.

I should note that, throughout this chapter, I will be concerned with two-party joint actions only. Joint actions involving larger groups are somewhat special in that, in those cases, it may suffice that a proper subset of all participants entertain a certain end. In this chapter, I will not be concerned with these intricacies.

Belief, knowledge, common knowledge, and belief of common belief. First, I will

leave the notion of belief unanalyzed. Second, in the context of this chapter, I only care about the difference between belief and knowledge insofar as knowledge is factive; knowledge entails *true* belief. Third, several agents commonly know a proposition *P* only if each knows that *P*, each knows that each knows that *P*, and so on *ad inf*. Furthermore, if an agent knows that a certain proposition *P* is common knowledge, then it is in fact common knowledge (Bonanno 1996). All this is straightforward. Common belief, however, is a bit more entangled. A proposition is commonly believed, if each agent believes that *P*, believes that each believes that *P*, and so on *ad inf.*; but unlike the case of common knowledge, if an agent believes that a proposition is commonly believed, then this does *not* entail that this proposition is commonly believed (Bonanno 1996). But apart from Bonanno's formal rendition, the idea is also independently intuitive. After all, an agent's belief that a proposition is commonly believed might just be false.

Note also that common knowledge (belief) merely *entails* such nested knowledge (belief); this is *not* to say that this nested structure *constitutes* common knowledge (belief). Famously, David Lewis (see Lewis 1969; Cubitt and Sugden 2003, 185) argued that common knowledge is *defined* in terms of symmetrically positioned reasoning on the part of various agents. Such symmetry is then said to *entail* nested knowledge. Whether Lewis was right need not concern us, because, by contraposition, a failure of such nested knowledge entails a failure of common knowledge.

Throughout this chapter, I will be conveniently using examples in which the

participants in a joint action have *actual* false beliefs about their co-participant's (and sometimes their own) intentions. This may seem like a departure from the philosophical literature. After all, philosophers in a broadly Lewisian tradition couch things in terms of mere *dispositional*, or *potential* beliefs (for a review see Paternotte (2011)). In this chapter, I'm concerned with the question of whether joint action requires that the *contents* of these beliefs (actual, dispositional, or potential) be true. Hence, I'm not concerned with the question of whether joint action requires common knowledge to be specified in terms of a particular epistemic modality. Therefore, we could phrase the examples presented here in terms of (say) dispositions to believe. To indicate what this would look like, consider the following rephrasing of **Lucky Jog**: Sarah and Bob both intend that they go jogging. Sarah has a dispositional belief that Bob would continue the jog even if it rained. Her dispositional belief about Bob, however, who would bail if it rained, is false. Fortunately, sunny weather prevails and they complete a happy jog. In this sense, I will argue that joint action sometimes requires false dispositional or potential beliefs.

Let me address one related preliminary worry. Paternotte (2011) observes that common *knowledge*, taken as face value, requires an excessively high epistemic standard and is, therefore, not of much use for anything. In most everyday cases, Paternotte argues, we don't have knowledge of other people's mental states, but, rather some type of probabilistically justified belief short of being knowledge. To see this, Paternotte has us imagine that a sentence is publicly uttered in the presence of person *A* and person *B*. This sentence, many would hold, is now common knowledge between us. But this is false, Paternotte opines: "There is no way for A to be sure that

B correctly heard the statement E" (see Paternotte 2011, 255). Absent such certainty, we could not have common *knowledge* of this event which is why Paternotte crafts a definition of common knowledge that employs lower standard of justification. Now, with this in mind, one might worry that these weakened notions of common knowledge already account for the possibility of error in our mutual belief attributions which is why the present chapter might seem superfluous. I think this criticism is incorrect for two reasons. First, the claim presented in this chapter is not merely that joint action is compatible with such false beliefs, but, rather, that it sometimes *requires* these beliefs to be false. Second, Paternotte is concerned with the fact that, in ordinary cases, a person's beliefs about what yet others believe *might have been false* (but are, in fact, true) which is why they do not amount to knowledge. For Paternotte this is reason enough to lower the justificatory standards for common knowledge. In the present chapter, I will defend the claim that, quite often, joint action requires that some of the beliefs about others' intentions are *in fact false.* This, however, is not Paternotte's concern. Let's examine an example to gain a clearer understanding of the intended contrast:

> **Table or Desk.** You and I are participating in a game show. Although we're positioned next to one another, neither of us can see what the respective other is doing. Through a loudspeaker the word "table" is uttered in a clear and comprehensible manner. We each have a scrap of paper and we are tasked with writing down the word that was uttered over the loudspeaker. We are given forced choice between the words "table" and "desk". If we both correctly write down the word that was uttered over the speaker we

> each receive 1000 dollars. If we both write down the same, but wrong, word we each receive 100 dollars. If we write down different words, we receive nothing.

According to philosophical orthodoxy, the fact that "table" was uttered over the loudspeaker creates common knowledge that this word was uttered. This piece of common knowledge, in turn, is important for coordinating on the "table" equilibrium. To see this, suppose that I know that you heard the word "table" but also thought that you think that I heard the word "desk". In this case, I would think that you will write down "desk" which is why I should write down "desk" trying to match your decision. Now, Paternotte's point, as I understand it, is that such coordination games do not require actual common knowledge that "table" was uttered, but, rather, a less demanding form of probabilistically justified belief. His point is not that coordination is possible if our beliefs about what the other person heard were, in fact, false. Cases in which such cooperative activities are possible while the beliefs about the others' intentions are *actually* false are much harder to come by. In this chapter, I explore these possibilities with regard to joint action.

Rationality. Let me add a disclaimer: I will discuss joint action under the assumption of common knowledge of rationality. This is a strong assumption in need of some justification. First, it is simply worth investigating whether fully rational agents could act jointly without common knowledge of some of the relevant intentions. Second, in the philosophical literature, common knowledge of all pertinent intentions is often added for the reason that rational agents would need it to engage in

joint action and related cooperative activities (e.g. Lewis 1969; Bratman 1987; Rubinstein 1989; Blomberg 2016). Arguing that common knowledge failure is often required to enable joint action involving ideally rational agents marks a natural extension of the extant literature. Third, if ideally rational agents do not need common knowledge of their intentions to successfully engage in joint action, then the idea that such knowledge is required for joint action is *thoroughly* undermined, because a common knowledge requirement is often added precisely for the reason that rational agents would need it. That said, exploring necessary condition for joint action without assuming common knowledge of rationality is a worthwhile endeavor in its own right. Unfortunately, it'll need to be left for another time.

3. Two types of assurance in joint action

In this section, I will argue that rationally intending a joint action requires that each participant enjoys some appropriate degree of assurance concerning the identity and robustness of the intentions of each. Thereafter, in the next section, I will argue that, in certain cases of lucky joint action, the success of joint action depends on these beliefs being false.

The first type of assurance (i.e. a belief that rationalizes action) concerns beliefs about the intended *ends* of each participant. The second type of assurance concerns beliefs about the intended *subplans* of each participant. I will discuss each type of assurance in turn.

Let's start with the following, rather uncontroversial, constraint on rational intending:

> **Rational Intention.** One ought not to intend what one believes to be impossible.[27]

In many cases, the absence of common knowledge of the relevant intentions will undermine **Rational Intention**. This is evidenced by the following vignette:

> **Cards.** We each intend that we build a house of cards. We each know that we intend that we do so. I, however, falsely believe that you falsely believe that I intend that we play Blackjack.

Given my belief that you believe that I intend that we play Blackjack, I will reason that you will not act on your intention that we build a house of cards; at least, that is, if you are rational. After all, the satisfaction of your intention depends on me enacting my part, which you believe I won't do. Analogously, the satisfaction of my intention depends on you enacting your part, which I now don't think you will do. This puts me in a position to believe that, at least if we are rational, I won't be able to satisfy my intention, which, in turn, makes the pursuit of this intention irrational. The moral, then, is this: common knowledge of the relevant intentions assures each participant that their intention in favor of the joint action is not, in fact, unsatisfiable. Common knowledge failure is then taken to undermine such assurance.

Above, I said that a failure of common knowledge will "in many" (but not in all) cases violate **Rational Intention**. Olle Blomberg (2016) has argued – I think convincingly – that, in some cases, a failure of common knowledge of these intentions

[27] e.g. Davidson 1978; Tuomela 1974, 133; M. E. Bratman 2013, 71; Grice 1957, 1989.

need not violate **Rational Intention**. He gives the following example:

> **Hector and Celia.** Hector and Celia are about to build a block tower. Each intends that they build a block tower, and each intends to do their bit of this joint performance. [...] Hector falsely believes that Celia falsely believes that he intends to cover the top face of each of her blocks rather than to do his bit of their joint performance. (Blomberg 2016, 318)

Blomberg argues that **Hector and Celia** is compatible with joint action, in part, because it is compatible with rationally intending the joint action (Blomberg 2016, 319). After all, Hector will think that it doesn't really matter whether Celia thinks that he merely intends to stack wooden cubes on top of her cubes. Stacking blocks on top of her blocks is exactly what he would need to do to act in favor of her intention to build a block tower. Hence, he shouldn't expect her to abandon her intention to build a block tower. Analogously, Hector should likewise believe that his own intention that they build a block tower can be satisfied.

Importantly, even in Blomberg's example the participants in a joint action need some kind of assuring beliefs concerning the intentions of each. His argument is simply that these beliefs need not be a correct representation of the intentions of each. To see this, consider an amended scenario in which Hector falsely believes that Celia intends to play 'race car' with the blocks. Surely, in this case, it would be quite irrational for him to intend to build a block tower with Celia (for the familiar reasons provided in the discussion of **Cards**). Blomberg's point is that the true intentions of each don't need to become common knowledge; the point is not that no assuring

belief structure needs to be in place. Hence, even in this example, each participant needs to be assured that the other's intentions are, in some sense, appropriately related[28] to their own intention. If Hector didn't believe that their intentions were so-related Hector would be irrational in retaining any of the intended subplans in favor of building a block tower.

Furthermore, in Blomberg's example, joint action is compatible with each knowing the respective other's interdependent end. Both do intend that they build a block tower. Common knowledge of this intention would, surely, not undermine joint action. Knowing the other's intentions isn't necessary but wouldn't hurt either. In the next section, I will discuss cases in which such knowledge would, in fact, undermine the possibility of joint action.

I will now turn to a second type of assurance in joint action that is closely related to Bratman's idea that, in joint action, the agents' subplans need to exhibit some (yet to be specified) degree of "mesh" (M. E. Bratman 2013, 54). Bratman gives an illustrative example: Suppose we each intend that we go to NYC.

> Your and my subplans can mesh even if they do not match. Perhaps your subplan specifies that we not go during rush hour, whereas mine leaves that issue open yet our sub-plans are co-realizable. Further, what is central to shared intention is that we intend that we proceed by way of sub-plans that mesh. This can be true even if, as we know, our sub-plans do not now mesh, so long as we each intend that in the end our activity proceed by way of a

[28] Due to space constraints, I have to let the notion of "appropriately related intentions" unanalyzed.

> so-lution to this problem. Nor need we each be willing to accept just any specifi-cation of activities of each that would suffice for the intended end. (M. E. Bratman, 2013, 54).

One of the ideas expressed in this example is that while subplans don't need to be fully specified, we nevertheless need not be willing "to accept just any specification". I'd now like to elaborate on this idea and formulate a necessary condition concerning the required level of specification for joint action.

The rough idea is this: In joint actions, agents need to be assured not only that the intended action is possible (as argued above), but also that the intentions of each will persist through a range of counterfactual situations; i.e. that these intentions are sufficiently robust.[29] To see this, consider the following vignette:

> **Canceled Jog.** Jane and Trevor both intend that they go jogging. It is common knowledge between them that they so intend. Jane, however, believes that Trevor will bail at the slightest sign of rain. Jane, on the other hand, doesn't care about the rain. In fact, she greatly enjoys the rain while jogging. There are clear signs of rain today; so Jane, thinking of Trevor's disposition to bail, abandons her intention that they go jogging.

Jane's decision seems perfectly rational. Given that Trevor will likely bail, it's simply too risky for her to set out to go jogging with him. Hence, for Jane it is not enough to

[29] To be sure, Bratman points to "three forms of persistence interdependence" of intention (Bratman 2013, 70ff) in joint action. Roughly these are: First, in many cases, my intention to do something with you often depends on you persistently wanting to do this thing with me. Second, we must both continue to believe that our success is realistically possible. Third, sometimes persistence of our intentions is grounded in moral obligation (e.g. due to a promise). Bratman's thoughts on intention persistence, as will become clear, are orthogonal to my discussion of believed mesh of intentions.

know that Trevor intends as she does. She needs further assurance that his intentions are somewhat robust, or, as Bratman would have it, that there is sufficient degree of mesh between their subplans. But how robust exactly?

When formulating a constraint that captures Jane's reasoning in **Canceled Jog** we have to be careful. We certainly don't want to stipulate a fixed credential threshold to indicate a participant's certainty that the intentions of each will be satisfied. To see this, consider the following example:

> **Risky Jog.** I promise you one million dollars if you go on a five-mile jog with my friend Bob. You know that Bob will bail at the slightest sign of rain. It's looking rainy right now and you believe that, most likely, Bob will bail before the five miles are completed. The payoff, however, is alluring; so you intend that you and Bob go on a five-mile jog.

In **Risky Jog**, you have a low degree of belief that your intention will be satisfied. However, given the particular settings of the case, it seems perfectly reasonable for you to intend as you do. Hence, we need a constraint weak enough to capture the reasoning behind **Risky Jog**. I propose the following provisional definition: If two agents, *A* and *B*, intend that they *J*, then:

> **Moderate Robustness (provisional).** For each agent, there exists a minimal set of circumstances, C_A (C_B), such that, if *A* (*B*) continues to intend that they *J*, then *A* (*B*) believes that it is commonly believed by all participants that they each intend that they *J* in all circumstances specified in C_A (C_B).

This is a mouth full. But an illustration will illuminate the mechanics of this principle. In **Canceled Jog**, Jane's set C_J contains rainy circumstances. She will continue to intend that they go jogging, only if she believes that she herself and Trevor would continue the jog in rainy circumstances. Hence, in **Canceled Jog**, she abandons her intention that they go jogging, because she does not believe that Trevor's intention would persist under rainy circumstances. In this case, it is false that each believes that all participants' intentions persist under rainy circumstances, because Jane herself doesn't believe that Trevor's intention would so persist.

C is a "minimal" set. This means that a participant would abandon her intention given any proper subset of *C*. Furthermore, an agent's set *C* is held rationally, if it maximizes the agent's utility. The thought is simple: Suppose Jane thinks that it will likely rain and the disutility she would experience from Trevor's bailing under rainy circumstances would be great. In this case, the expected utility associated with going on a jog with Trevor might just be too low to justify maintaining and acting on her intention.

Let's, next, be precise about the exact content of the belief of common belief specified in **Moderate Robustness (provisional)**. The added precision will lead to a slight amendment of this principle. Let's again use **Canceled Jog** as an example. Jane's intention that they go jogging is predicated on her belief that Trevor will jog even if it rains. Rainy circumstances are part of her set C_J. Does she have to believe that rainy circumstances are also part of Trevor's set C_T? No, because although Trevor might be happy continuing the jog under conditions of rain, he might be equally

happy to seek shelter should Jane wish to do so. Rainy circumstances might not, as it were, be part of his *minimal* set C_T; i.e. his intention that they go jogging might not depend on him thinking that she will continue the jog in rainy circumstances. Now, intuitively, it only matters to Jane *that* Trevor won't bail if it rains; it does not matter to her whether running in the rain is part of his minimal set C_T. Instead, Jane needs to believe that Trevor believes that she shares all the elements specified in the set C_T indexed to *him*.

Furthermore, Jane needs to believe that Trevor believes that Jane believes that he shares the elements of C_J indexed to *herself*. To see this, consider the following line of reasoning: if Trevor thought that Jane believed that he didn't share the elements in her set C_J then he would be in a position to reason that she will abandon her intention that they go jogging. This, in turn, would rationalize his abandoning the intention that they go jogging. The upshot, then, is this: Trevor and Jane are required to entertain the nested type of beliefs typical of common belief; however, the contents of these nested beliefs alternate from one level to the next in the way indicated above. Based on these considerations we should reformulate the above principle as follows: If two agents, A and B, intend that they J, then:

> **Moderate Robustness.** For each agent, there exists a minimal set of circumstances, C_A (C_B), such that, if A (B) continues to intend that they J, then:
>
> *(i)* A (B) believes that they each intend that they J in all circumstances specified in C_A (C_B).

(ii) *A* (*B*) believes that *B* (*A*) believes that they each intend that they *J* in all circumstances specified in $C_{\boldsymbol{B}}$ (C_A). (The crucial point is that the indexes "$_B$" and "$_A$" are reversed as compared to *(i)*. More generally, these indexes reverse at each level relative to the previous one.)

(iii) *A* (*B*) believes that *B* (*A*) believes that *A* (*B*) believes that they each intend that they *J* in all circumstances specified in $C_{\boldsymbol{A}}$ (C_B) and so on *ad inf.*

This concludes my rendition of assurance in joint action. First, the participants need to be assured that the satisfaction of their intentions is possible. Second, the participants need to be assured that the relevant intentions favor the joint action in a robust fashion.

4. Intentions in the context of assurance

In this section, I will put to work what we have learned so far and argue that, given each participant's enjoyment of the kind of assurance set out above, acting jointly sometimes depends on misrepresenting the intentions of one's co-participant. To start seeing this, re-consider **Lucky Jog**:

Lucky Jog. Sarah and Bob both intend that they go jogging. Sarah believes that Bob would continue the jog even if it rained. This is important for her! Her intention that they go jogging is conditional on her belief that he wouldn't bail if it rained. Her belief about Bob, however, who would bail if

it rained, is false. Fortunately, sunny weather prevails and they complete a happy jog. As it happens, they got lucky.

Consider also another example with larger scope:

Lucky Marriage. Ian and Mia stand on the altar each vowing: "I promise to be true to you in good times and in bad, in sickness and in health." Both Ian and Mia intend this to be an honest expression of their conviction; they are, as it were, sure that they will be able to keep this promise. So they get married. By all measures, their marriage turns out to be truly wonderful. It lasts for 50 happy years until they both pass away. The "bad" times, in which, as they promised, they would stay together never came. There was no sickness, no temptation and the like. Had such bad times come around, Mia would not have stood by Ian. The same is true of Ian. However, had it been known to them that their marriage could not endure such bad times, they would have never gotten married in the first place. In fact, had it been known to them that they themselves could not keep this promise, they would not have gotten married either. Their belief that the respective other had robust intentions to stay together was crucial to them, despite the fact that, under pressure, it would not have held up. They both got lucky![30]

Sarah, as well as Ian and Mia have false beliefs about the robustness of the respective other's as well as their own subplans. Each misrepresents the robustness of their co-

[30] It might be objected that a marriage somehow isn't a joint *action* (but rather a *project*). The example is included mainly to show how ubiquitous the falsity of such robustness related beliefs really is.

participant's as well as their own intentions. Nevertheless, the actions described in these vignettes are perfectly joint for the following reasons: First, it just seems intuitive that these are cases of joint action. Should we pay attention to such intuitions when defining joint action, or should we instead keep our focus on independent theoretical considerations? I think we should, in fact, pay attention to our considered intuitions for the following reasons. First, there is no agreement about the theoretical role of joint action. Butterfill (2011), for instance, argues that joint action plays a role in *enabling* mindreading in infants. Joint action should, therefore, not presuppose mindreading which is why Butterfill opts for a sparse definition of joint action. However, whether (and the extent to which) mindreading needs *enabling* is, in fact, contentious and heavily debated in the literature (for a review consult Carruthers (2013)). Some have argued that mindreading is a thoroughly innate capacity and can be traced experimentally even to infants as young as 6 months. It seems to me that the theoretical role of joint action is insufficiently understood which is why basing one's definition of joint action on the theoretical role proves difficult.[31] Future research of the sort initiated by researchers such as Butterfill (2012) and, relatedly, Vesper *et al*. (2010) will show whether findings in, say, developmental psychology can help ground definitions of joint action.

That said, one uncontroversial theoretical desideratum of joint action theory is that it is a basic and ubiquitous form of social interaction (see introduction). This

[31] Second, theory-driven definitions of joint action need to respect our considered intuitions on pains of not changing the subject. A case in point is an exchange between Butterfill (2011) and Blomberg (2015) who criticizes Butterfill's (2012) minimal account of joint action precisely because it overextends to cases that we would not intuitively label to be cases of joint action. This is not the point to discuss their exchange; however, we should note that theory-driven definitions of central philosophical concepts often run the risk of changing the subject by abandoning the initial target phenomenon. Explicating the precise role joint action plays in broader social scientific concerns is, of course, an important concern.

assessment would have to be revised if it were denied that **Lucky Jog** and **Lucky Marriage** are joint actions. After all, there is nothing at all contrived about these cases. Jointly acting agents are regularly mistaken about the degree of robustness of their co-participants' intentions. In some cases, the problematic counterfactuals do materialize, which, in turn, leads to a breakdown of the joint action. However, in many other cases, in which the relevant counterfactuals do not materialize, this leads to the pursuit of perfectly normal joint actions.

A critic might object that joint action marks, in some sense, an ideal way of acting together; a way of acting together that is particularly *safe* and can't, therefore, come about through lucky circumstances. I think this criticism is misguided. Quite frankly, joint action does not mark an ideally safe way of acting together. Rather, – and the philosophical literature concurs – joint action is a basic and ubiquitous type of social action (see above). It is, as it were, the *actual base* of sociality, and not an ideal to live up to. There are in fact countless ways to make joint actions more safe and reliable (e.g. through third-party enforced contracts, explicit promising, mutually known expectations, or habitual action (e.g. Michael and Pacherie, 2014) that go way beyond what is minimally required for joint action to occur.

Therefore, giving up on the idea that joint action is ubiquitous and socially basic is a high theoretical cost. After all, philosophical interest in joint action is, as I showed at the beginning of this chapter, stoked, at least in part, by the conviction that joint action is a fundamental way of acting together.

The lesson from this discussion, then, is this: In the context of joint action, we

should require that all participants have some appropriate degree of assurance concerning the range of subplans intended by each participant. We should, however, not require that these beliefs be true. Furthermore, in **Lucky Jog** and **Lucky Marriage**, both pairs of agents engage in join action, yet knowledge of the range of subplans intended by each would readily undermine the joint action.

Let me now apply this insight to a second set of cases. In these cases, the participants' intended interdependent ends are incompatible and knowledge of these ends would undermine the joint action. To see this more clearly, re-consider **Forking Trip.** In this case, going to Philadelphia requires that we each have false beliefs about the respective other's end. If we knew that our ends didn't match, we would abandon our trip altogether and our jointly going to Philadelphia would be undermined. Our jointly going to Philadelphia depends on our false beliefs about the intended interdependent ends of each. The intuition, however, is that we jointly go to Philadelphia. Given the above analysis of assurance, we're now in a position to defend this intuition on principled grounds.

To be very clear, the claim is *not* that **Forking Trip** is entirely devoid of matching interdependent intentions. Surely, in this example, both of us have matching interdependent subplans; namely that we go to Philadelphia; and this subplan is itself an interdependent intention with a "that we φ"-type content. Rather, merely our intended interdependent *ends* are incompatible and not jointly realizable. Why should we believe that **Forking Trip** presents a case of joint action?

First, our interaction, while on the way to Philadelphia, is, in all important

behavioral respects, indistinguishable from the case in which we both intend, as an end, that we go to Philadelphia. In both cases, we appropriately support one another, and our actions are equally interdependent.

Second, a critic might further object that the impossibility to co-realize both ends is problematic for a different reason; namely, that it renders the ends irrational. This objection is, again, misguided. According to the intuitive principle **Rational Intention**, rationally intending does not require that it be possible to satisfy the intention. Rather, mere *belief* that the satisfaction is possible is required. A stronger and, arguably, more controversial principle of rationality is required to render the relevant ends in **Forking Trip** irrational.

Third, the critic might go on objecting that the relevant beliefs are not robust enough, because joint action requires, as Kutz (2000) puts it, that "no one would modify his or her plans in virtue of disclosure." In **Forking Trip**, however, disclose of our ends would undermine the joint action. This point is no doubt true; but it overextends. In **Lucky Jog** and **Lucky Marriage**, the joint action would not survive the revelation of the actual robustness of each participant's intention. However, at least in those cases, this does not sanction the judgment that these actions are not truly joint actions. The critic would have to further argue that the actual ends are special in some sense, such that the joint action should survive the revelation of the ends, but not their subplans. This move, however, seems unmotivated, and ought to be rejected for the sake of the theory's consistency and simplicity.

Let me present one pressing objection in detail. Ben Laurence (2011) and Hans

Bernhard Schmid (2016) have recently employed and expanded the Anscombian idea that acting intentionally requires, quite generally, some form of teleological rationalization (Laurence 2011; Schmid 2016). Thus, intentionally acting agents can answer the question "Why?" they act as they do. To illustrate, suppose you fill up a kettle with water in order to make tea. When asked "Why should you want to fill it up?'" (Laurence 2011, 278) you might answer "Oh, because I want to make tea" (Laurence 2011, 278). Intentional action is said to depend on such rationalizations. Similarly, intentional *joint* agency is, then, likewise said to depend on the availability of such rationalizations. What is more, each participant's contribution to a joint action needs to be rationalized with reference to a *joint* (i.e. collective) action. To illustrate further, suppose a band of robbers are robbing a bank. The various robbers have different tasks. One robber's job is to crack open the safe. If their robbing the bank is indeed a joint activity then the robbers' individual actions (e.g. cracking open the safe) are rationalizable with regard to the joint end of robbing the bank. If asked, "Why is he cracking the safe?" the answer "because the band of robbers is knocking over Mellon bank" (Laurence 2011, 278) is appropriate. Laurence further requires that in joint actions such rationalizations should be *the same for each participant* (see Laurence 2011, 282). Hence, the explanation "because the band of robbers is knocking over Mellon bank" marks the teleological endpoint of their joint action that rationalizes each robber's contribution.

This line of reasoning generates an apparent objection against the claim presented in this chapter. Presumably, in **Forking Trip**, each of us will rationalize their actions differently. When asked "why are you on this bus?" I will answer "because we are

going to NYC", citing *my* interdependent end; you, on the other hand, will answer "because we are going to Ocean City", citing *your* interdependent end. But each of the rationalization that we provide are false. We're neither jointly going to NYC, nor are we jointly going to Ocean City. Our joint action merely extends as far as one of our subplans. Hence, the teleological rationalization that each of us provides is false.

To answer this objection, we should first understand why we should accept this particular rationalization criterion in the first place. Schmid stresses that intentional agency requires that agents have available some rationalization for their actions. He explains that "it has to be apparent *to us* what it is we're doing, or else intentional action breaks down" (Schmid 2016, 52). To see this more clearly, he provides the following example for such a breakdown in agency:

> **Fridge**. "Imagine that during a short break after some hours of intense work on a paper at your desk, still thinking about your paper, you find yourself in the kitchen, opening the fridge, not knowing what it is you're doing there. Perhaps your cluelessness does not run all the way down to your present bodily movements—you know perfectly well that you're opening the fridge—but you have no idea as to the question of why you're doing it. Were you about to get something from there, or put something back? You still feel utterly lost and rather stupid for a moment, and intentional action has broken down […]" (Schmid 2016, 52).

Individual agency, Schmid argues, requires that an agent *can answer* the relevant "Why?" question (not just that the question is answer*able*) by citing her end. Joint

agency, as Schmid and Laurence suspect, requires that participants can answer the "Why?" question by citing their joint end.

Now, I take it that, in **Forking Trip,** the mere *availability* requirement of such a rationalization is satisfied**.** After all, in said vignette both of us will rationalize their actions with regard to an intended joint action. If you are asked "Why are you on this bus?" you will answer "because we're going to Ocean City together"; likewise, I will answer "because we're going to NYC together." Hence, although false, we do have a rationalization at hand.

Should we add the further requirement that each of our answers to the "Why?" question be *true*? The answer, I think, is not obviously "yes." The simple reason is that agents can be mistaken about the particular explanations they invoke to rationalize their actions. Yet such confabulation does not obviously seem to render their actions unintentional (unlike the way the actions in **Fridge** seem unintentional and devoid of agency). Carruthers (2011, 342) and Wegner (2002) discuss experimental cases in which subjects carry out an instruction that was given to them under hypnosis and who "will often confabulate an explanation for their action citing some or other particular intention" (Carruthers 2011, 342). For instance, subjects will follow the instruction "when I see the book on the table, I shall place it on the shelf." When later asked why they placed the book on the shelf, subjects confabulate an intention such as that they intended to tidy the room. Similarly, Schmid contends that our intentions are not always transparent to ourselves. He invokes the example of two friends quarreling all evening and only later realizing that the intention behind the

quarrel was to break up the friendship (Schmid 2016, 57).

Although we should agree that agents should be able to give a rationalization for their intentional actions, it seems at best controversial that intentional agency should depend on the *truth* of those avowed rationalizations. Now, the rationalizations that the agents in **Forking Trip** would provide are false. When I answer the question "Why are you on this bus?" by saying "because we're going to NYC", I'm saying something false; however, as we've seen we shouldn't require such rationalizations to be true – neither in the context of individual agency, nor in the context of joint agency.

Summing up, we should agree that, in the context of joint action, there should be a description under which we act jointly. This description may well correspond to a subplan of ours. In **Forking Trip**, this description is "jointly going to Philadelphia". We can further concur with Schmid's claim that intentional agency requires that an agent be able to have available an answer to the relevant "Why?" question. After all, in **Fridge** intentional agency really seems to have broken down; hence intentional agency seems to depend on the availability of such rationalizations. We can also concede that, in the context of joint action, the answer to this "Why?" question should refer to a joint action (e.g. "I φ because *we* ψ"). However, we don't need to concede that the answers to such "Why?" questions necessarily be veridical. Bar such a veridicality requirement, the Anscombian approach to joint action does not undermine the jointness of **Forking Trip**. For these reasons, I conclude that there are no reasons to doubt that Forking trip is an any less genuine case of joint action than

the ones put forward by the traditional accounts.

Lastly, a critic might object that, in **Forking Trip**, both of our ends could not, in fact, be co-realized; however, the objection continues, we should require that, in the context of joint action, the ends of each must be mutually co-realizable. Note that this objection is close to being a brute denial of the idea that **Forking Trip** is a joint action; after all, one of the central claims of this chapter is that, in joint actions, the intended ends don't need to be co-realizable. As such, the objection is much more a mere assertion of a clash of intuitions than it is a structured objection. A *mere* clash of intuition produces a stand-off. However, I think by now we have gathered extensive evidence supporting the idea that this example does present a case of joint action. We saw that this example is compatible with an intuitive way to spell out rational intending, it is a natural extension of our assumption that joint actions are socially basic, and the behaviors specified in this example are behaviorally indistinguishable from more mundane cases of joint action.

So far, we have looked at examples of lucky joint action in which incompatible intentions remain undisclosed to the participants for the duration of the joint action. We may further ask whether joint action depends on the non-disclosure of the pertinent incompatible intentions even after completion of the action. The answer, I think, is that it does not. But let's first explicate why one might think that it does. Reconsider, **Forking Trip**; suppose that, having arrived in Philadelphia, we find out that we are on our way to different cities. In hindsight we might be inclined to judge that, really, we didn't jointly go to Philadelphia after all. Concomitantly, we may

want to say that our jointly going to Philadelphia depends on our not finding out that our intentions were incompatible all along. I think this assessment is wrong. To see this, we should, once again, see the parallel to **Lucky Jog**. Suppose we come home from a nice jog. As theorists, I have argued that, at this point we are entitled to judge that Sarah and Bob jointly went for a jog. If lucky joint action depended on non-disclosure of the incompatible intentions, this judgement would be inappropriate. Rather, we would only be entitled to judge that Sarah and Bob went on a jog together *provided that they won't find out about the discrepancy in intention later on*. This, I think, is implausible. After we have come home from a nice jog, the question whether we went on a jog together is settled. Similarly, consider the case in which, having arrived in Philadelphia, we both get urgent calls from home and the both of us have to rush back to Baltimore. At this point it seems to be settled – given the above analysis is right that is – that we jointly went to Philadelphia. If Lucky Joint action were dependent on persisting non-disclosure of the discrepant intentions, we would have to hold off on this judgement, because our discrepant attitudes may be disclosed at a later point in time.

5. Conclusion

In this chapter, I've argued that joint actions can be "lucky". In these cases, jointly acting agents are mistaken about their co-participant's subplans and interdependent ends. Furthermore, in the relevant cases, the joint action *depends* on these mistaken beliefs; common knowledge of the relevant intentions would undermine the possibility for joint action. Hence, the analysis of "lucky joint actions", as I called

them, shows the extent to which common knowledge can be harmful to joint action.

Chapter 4: Coordination Through Precedent Without Common Inductive Standards

1. *Introduction*

Consider the following case:

> **Fast Food.** You and I want to meet for lunch. We have two options: McDonald's or Wendy's. We don't care where we'll have lunch as long as we'll have lunch together.

In situations such as this one, each of us is trying to predict what the other is going to do in an attempt to match our choices. These predictions can be justified in various ways. Maybe we have antecedently agreed or promised to each other that we would go to McDonald's. But suppose that we have no explicit agreement to go on, but that, instead, we have always gone to McDonald's in the past. In this case, you may rely on precedent and simply reason as follows: "Since that's where we've always gone, that's where she'll go again". For ease of presentation, let's call the proposition "We've always gone to McDonald's" proposition A, and the proposition "You (or I respectively) will go to McDonald's" proposition $x_{y/I}$. We can, then, represent the evidential relation between A and $x_{y/I}$ as follows:

(1) $A \rightsquigarrow x_{y/I}$

The squiggly arrow indicates that the inference is defeasible; i.e. valid only by default; the fact that we've always gone to McDonald's does, of course, not

necessitate that this is what we shall do.[32] In predicting your behavior, I may use what's stated in **(1)**. Let $K_{I/y}$ denote "I/you know that ...". We may, then, represent my reasoning process as follows:

(2) $K_I(x_y)$ because $K_I(A)$ and $K_I(A \rightsquigarrow x_y)$

I know that you'll go to McDonald's, because I know that this is what we've always done, and because I know that people tend to continue to act as they have in the past.

Although, these claims are couched in terms of knowledge (instead of, say, certainty, belief, degrees of belief, or reason to believe) this should be taken more as a convention rather than a substantive commitment. As Harvey Lederman (2018) stresses, in debates about coordination the term "common knowledge" is often used as a catch-all term for "common knowledge, common belief, and common certainty". In this chapter, knowledge will often be important only insofar as it entails belief; and many of the examples that follow will be couched in terms of (common) belief.

Now, it has been said that, at least in the context of coordination games, the line of reasoning presented in **(3)** is incomplete. Famously, David Lewis has argued that predicting another agent's behavior using precedent as a source of evidence depends on the "mutual ascription of some common inductive standards […]." (Lewis 1969, 56f). Similarly, Sugden and Cubitt, analyzing Lewis on convention, require that the players "have reason to believe that, in particular relevant respects, they have common background information and common inductive standards" (Cubitt and

[32] For this reason Cubitt and Sugden (2003) say that *A* gives a "reason to believe" that *x* is the case. Alternatively, using Lewis' preferred language, we can say that *A* "indicates" that *x* is true.

Sugden, 2003, 185). A common inductive standard is a piece of knowledge (belief, or reason to belief) that a particular salient feature (e.g. precedent) is a "projectible regularity" (Cubitt and Sugden 2003, 198) such that "each person must have *reason to believe that each other person shares his own standards* about what can be inferred inductively from A" (my italics) (Cubitt and Sugden 2003, 198). Extending this thought, Christina Bicchieri has suggested that, to select an equilibrium in a coordination game, "we must introduce some salience criterion of choice, *and common knowledge thereof* […]. Salience may be provided by precedent [...]" (my italics) (Bicchieri 2006, 36). Similarly, Cyril Hédoin has argued that the players need "*common knowledge* […] that they share the same reasoning modes" (Hédoin 2014, 380). In summary, the thought is that simply *having* the same standards of inference is not enough for successful coordination; rather, the players need additional beliefs or knowledge that the respective others are disposed to reason similarly.

To see the putative importance of a common inductive standard we should examine a case in which it is missing: Reconsider **Fast Food** and suppose for a minute that I am initially inclined to predict your behavior using precedent; suppose also that, this time around, I think that you falsely believe that I intend to *defy* precedent predicting that I have decided to go to Wendy's instead. Of course, in this case, it would seem quite silly to stick to my guns and use precedent as my inductive standard. I know that you are going to try to match what you believe I will do and, thus, go to Wendy's. If my goal is to coordinate with you, then I should likewise go to Wendy's. Hence, in predicting your behavior, I ought not only to know what we've done in the past, but, additionally, that you are also inclined to predict that I will go to

McDonald's using precedent as a standard of inference.

Adding a common inductive standard to formulation **(2)** we obtain:

(3) $K_I(x_y)$ because $K_I(A)$ and $K_I(A \rightsquigarrow x_y)$ and, as one precondition, $K_I(K_y(x_I)$ because $Ky(A)$ and $K_y(A \rightsquigarrow x_y))$. [33]

This statement reads as follows: I know that you will go to McDonald's in the future, (a) because I know that we've gone to McDonald's in the past, (b) because I know that people tend to act in the future as they have in the past, and (c) because I know that you think that I will go to McDonald's for the same reasons (i.e. because you know that we've done so in the past, and because you know that people tend to act in the future as they have in the past).

In this chapter, I will argue that the common inductive standard requirement expressed in **(3)** is implausible for two reasons. First, it implies a higher-order belief about what the other player thinks oneself will do. However, predicting another player's behavior based on precedent is incompatible with the presence of such higher-order beliefs. The following argument, which I will explicate in this chapter, shows this:

P1 – Interdependence. In a two-player pure coordination game between

[33] Vanderschraaf and Sillari (2013) formulate this idea in terms of "symmetric reasoning". Symmetric reasoners know that an inference that the agent herself can draw can also be drawn by another agent. Formally, this idea is expressed as follows: [$\mathbf{K}_i(A) \Rightarrow \mathbf{K}_i(E)$ and $\mathbf{K}_i(A) \Rightarrow \mathbf{K}_i\mathbf{K}_j(A')$, then $\mathbf{K}_i(A')] \Rightarrow \mathbf{K}_i\mathbf{K}_j(E)$; i.e. if an agent can infer E from A, and the agent can infer that the other agent knows A, then she is also in a position to know that the other agent knows E. The definiens says that for each agent i, if i can infer from A' that E is the case and that everyone knows that A' is the case, then i can also infer that everyone knows that E is the case.

player A[34] and player B, A's first-order prediction about what she takes B to choose provides sufficient reason for her own rational strategy choice.

P2 – Double Justification. A's belief about what she takes B to choose can be justified by appeal to (a.) precedent, or (b.) higher-order beliefs suitably characterized (e.g. what A believes B believes that A will choose).

P3 – Higher-Order Defeat. In pure coordination games, justifications based on higher-order beliefs always defeat justifications based on precedent.

—

C – Absence Precondition. Precedent can justify A's belief about B's strategy choice only if A has no second-order belief about what B thinks A will choose (e.g. beliefs about what B believes A will choose).

The argument makes explicit an evidential relation between first-order behavioral predictions, precedent, and second-order beliefs. As the chapter unfolds, we will see that the argument generalizes for second, third, and, ultimately, *n*th level behavioral predictions.

The second reason rendering **(3)** implausible is that, sometimes, we want to *explain* common knowledge that a certain inductive standard is used by appeal to the independent predictive power of precedent-based inferences. If these inferences were to presuppose such common knowledge, then these explanations would turn out to be

[34] The argument is stated from player A's perspective. This is just to avoid clutter. The reader should fill in the exact same argument as given from B's perspective.

circular.

Before continuing further, I should add a caveat. The "common inductive standards" requirement depends on various idealizing assumptions. For instance, the idea that reasoning from precedent depends on *infinitely many* nested higher-order beliefs about the inductive standard may depend on the assumption that the players are unbounded reasoners. Furthermore, the idea that a player may only base her predictions of the other's behavior on precedent provided that the other does so as well, depends, at least on the face of it, on the idea that the player is rational, and, for further iterations, on the idea that their rationality is common knowledge.

Many have found it implausible that successful and reliable coordination should be premised on infinitely complex higher-order beliefs. The remedy has usually been to relax at least one of the various idealizing assumptions. Harvey Lederman (2017) has argued that coordination can be facilitated by letting go of the assumption that the players' rationality is common knowledge. Players can be rational, but they need not be smug; i.e. they might not know that they are rational. Others (e.g. Kneeland 2012) have argued that bounded reasoners can coordinate without common knowledge. Yet others (e.g. Skyrms 2004) have explored coordination in entirely non-strategic contexts.

In this chapter, I won't follow these approaches and, thus, keep all rationality assumptions. These are strong assumptions and the decision to keep them needs a bit of justification. First, it is simply worth investigating whether fully rational agents would need common inductive standards to coordinate. Second, in the philosophical

literature, such common inductive standards are often added because rational agents would need them to coordinate their actions. Hence, a discussion of these standards under these idealized circumstances marks a natural extension of the extant literature. Third, the reason these assumptions are often dropped is to lend empirical validity to a particular model. Actual agents, it is sometimes argued, simply aren't unbounded reasoners and don't commonly know that they are rational. In this chapter, my primary concern is not empirical validity, but, rather, a principled investigation into the tension between reasoning from precedent and higher-order expectations. Lastly, any successful theory of coordination should hold up under idealized circumstances, as it would be quite surprising if successful coordination were to *require* cognitive limitations.

In the early days of research on coordination (e.g. Rubinstein 1989) and cognate cooperative activities such as joint action (e.g. Bratman 1989), and conventional behavior (e.g. Lewis 1969), common knowledge that each participant will choose a particular strategy was seen as a requirement. This sentiment finds its most rigorous expression in Rubinstein's *Coordinated Attack* and *Electronic Mail* games. Subsequently, many had noticed that such common knowledge requirements depend on various idealizing assumptions (see above) and that relaxing these assumptions may render these requirements unnecessary. In each case, it is nevertheless argued that common knowledge is, although not necessary for coordination, always compatible with it. This idea has recently been challenged. Common knowledge requirements can, at times, be *harmful* to coordination (see Lederman 2017; Schönherr 2018). The present chapter extends this nascent line of research, arguing

that solving a coordination game relying on precedent is in tension with common knowledge that this standard is used.

In the first part of this chapter (section 2 – 4), I will show why the common inductive standard requirement expressed in **(3)** is implausible. In the second part (section 5), I will provide an alternative. More concretely, in the next section, I shall detail **P1** and **P2**. In section three, I will explicate **P3**, which will put us in a position to see why **C** is true, and, thus, why **(3)** is false. In section four, I will show how common inductive standards obscure the explanatory relation between precedent as a plausible rule of inference, and common knowledge that this inference rule is used by the players. Lastly, in section 5, I will sketch a positive picture describing how coordinating agents should think of one another in the relevant situations. Put coarsely, predicting other players' actions using precedent presupposes a form of mutual belief *suspension* about the inductive standard used by one's co-participants.

2. Interdependence and double justification

The games I will be talking about are two-player, conflict-free, pure coordination games; i.e. games with multiple strict Nash equilibria[35] in which one player's gain does not require the other player's sacrifice. Such games can be represented by the following matrix[36]:

[35] A Nash equilibrium is a set of strategies such that no individual has an incentive to change her choice given the choices of the others.

[36] This matrix should be read as follows: The labels 'Player I', and 'Player II' represent the players. The labels "X", and "Y" represent the players strategies. The numbers represent the players' utilities. These utilities are a function of the players' strategies. In this chapter, I will be making use of players' *pure* strategies; i.e. a player's decision to play a strategy with probability 1. The expected utility of choosing a particular strategy is the utility associated with this choice given the expected pure strategy of the other player(s).

		Player 1	
		X	Y
Player 2	X	1,1	0,0
	Y	0,0	1,1

Figure: 1

In this game, players have to solve the *equilibrium selection problem*. There are two relevant pure[37] equilibria, {X,X} and {Y,Y}, and the players have to figure out a way to settle on one of them. Ultimately, each player is trying to match what she takes the other player to choose, which is why each player's choice depends only on estimates (beliefs, credences, or knowledge) about the other player's choice. This is all I shall say in defense and illustration of the **Interdependence** premise.

Let's move on to the second premise. Beliefs about the other player's choice can be justified in several ways. In this chapter, I will focus on only two sources of evidence: Higher-order beliefs and precedent (and, concomitantly, the combination of both). Let's start with precedent as a source of evidence.

Predicting behavior using precedent means inferring future behavior based on a past behavioral regularity. The validity of such reasoning can perhaps be explained by the fact that "we may tend to repeat the action that succeeded before if we have no strong reason to do otherwise" (Lewis 1969, 36). To see how precedent might do this, consider first an example from David Lewis: "I know very well that I have often seen

[37] Mixed equilibria won't matter for our purposes.

cars driven in the United States, and almost always they were on the right. [...] Given a regularity in past cases, we may reasonably extrapolate it into the (near) future" (Lewis 1969, 41). Many (Sillari 2008; Lewis 1969; Sugden 2015; Bicchieri 2006) acknowledge that precedent can provide coordinating agents with evidence despite the fact that there is no theory with regard to what it is that makes a particular feature salient. Past and future actions are never alike in all, but, rather, merely in some respects. Reasoning from precedent is therefore dependent on certain salient features of actions. This, however, does not detract from the fact that precedent is real; in any case, in this chapter I will assume that it is. Lastly, precedent-based reasoning is defeasible; it is a mere "last resort [for the players], when they [the players] have no stronger ground for choice" (Lewis 1969, 35). This idea was represented above by the squiggly arrow. Higher-order beliefs, as I will argue this chapter, turn out to be a "stronger ground for choice".

Let's move to higher-order beliefs as a source of evidence. To start seeing how higher-order beliefs can guide predictions of the other's behavior, consider the following vignette:

> **Fast Food 2.** You and I want to meet for lunch. We have two options: McDonald's or Wendy's. We don't care where we'll have lunch as long as we'll have lunch together.
>
> I learn that a source, who you believe to be infallible, told you that I will go to Wendy's this time around.

For ease of understanding, we can depict my epistemic situation as follows:

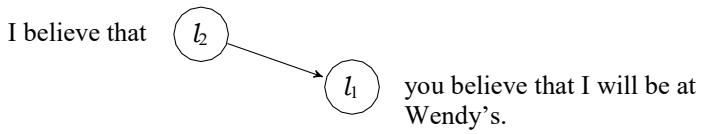

Figure 2

The indexed letter l_n indicates the depth of reasoning (e.g. l_2 represents a second-order belief). Note also that the arrow simply indicates that direction of the nesting of beliefs; it does not indicate a rule of inference of any sort. I reason as follows: Since you believe that I will go to Wendy's, you will intend to match my decision. Thus, I expect you to go to Wendy's, which is why the rational decision on my part is to go to Wendy's. My second-order belief about you justifies my first-order belief about where you will go. My first-order belief, in turn, settles my decision.

Such second-order beliefs can, in turn, be justified by a third-order belief as the following figure illustrates:

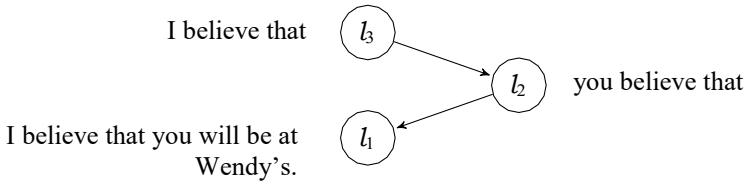

Figure 3

Of course, third-order beliefs can be justified by fourth-order beliefs and so on *ad inf.* Thus, the following general picture emerges: Any level-*n* belief about a player's choice can be justified, presumably, in various ways. One particular way to justify a level-*n* belief is in terms of a level $n + 1$ belief. If each belief is indeed justified in this way, the result will be an infinite hierarchy of justifying beliefs. Such an infinite

hierarchy of beliefs marks the crucial characterization (not a definition[38]) of common belief relevant for our purposes. Furthermore, if these beliefs are true, then this amounts to a characterization of common *knowledge*.

Although discussions of coordination games make frequent reference to common knowledge simpliciter that each will choose their part of a particular equilibrium, we should instead rely on the slightly amended notion of common *reciprocal* knowledge (and belief) which is gleaned from Robert Sugden (2015). The problem with good old common knowledge is its reflexivity; i.e. if *p* is common knowledge among individuals in a population *N,* then each individual in *N* knows that *p*, knows that she *herself* knows, etc. In coordination games such as **Fast Food**, however, each player is just concerned with what she believes *the others* are going to do, what others take yet others do and so on *ad inf*. As a reminder, this type of reasoning is illustrated in *Figure 2* and *3*. The Sugden-inspired notion of 'common reciprocal knowledge' captures this idea by taking reflexivity out of the definition. A group of players have common reciprocal knowledge (belief) that *p* is true, if, and only if, for all players *i* and *j* in *N,* where $i \neq j$, *i* knows (believes) that *p* holds for *j*; all individuals *i, j,* and *k* in *N,* where $i \neq j$ and $j \neq k$, *i* knows (believes) that *j* knows (believes) that *p* holds for *k*, and so on *ad inf.*.

Lastly, we should note that both types of justification can be combined; e.g. third-

[38] Although common knowledge (belief) can be *characterized* by an infinite hierarchy of (actual, potential, or dispositional) nested higher-order true beliefs, it should be noted that this is really just a *characterization*; not a definition. Definitions of common knowledge have, for instance, been given in terms of public events, or inference patterns between symmetric reasoners (for an overview consult Vanderschraaf 2014). These definitions, however, need not concern us, because, although common knowledge is not defined in terms of iterated beliefs, it nevertheless entails these iterations. Hence, by contraposition, a failure of such iterated knowledge likewise entails a breakdown of common knowledge.

order predictions can justify second-order beliefs that can, in turn, justify precedent. Here is a figure illustrating this thought:

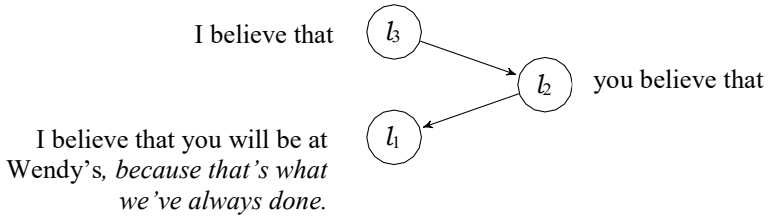

Figure 4

Similarly, we should say that each level *n* belief can, in principle, be justified with regard to a level *n+1* belief, or, alternatively, precedent.

3. Higher-order defeat

Precedent can justify predictions about the other player's actions only in the absence of any reciprocal belief that would likewise justify or undermine this prediction; this is because, as premise three of the above argument states, justifications based on higher-order beliefs defeat precedent-based justifications. Let's add some color and precision to this idea.

Let RB_n stand for any such *n*th level reciprocal belief. Let *Precedent* stand for a piece of precedent-based evidence bearing on this belief. Lastly, let RB_{n+1} stand for the reciprocal belief justifying RB_n. To illustrate, consider the following propositions:

RB_1 – I believe that you will go to Wendy's.

Precedent – In that past, we have always gone to McDonald's.

RB_2 – I believe that you believe that I will go to Wendy's.

The idea, then, is that RB_1 can neither be justified nor undermined by *Precedent* given the presence of RB_2. The reason is simple: I expect you to think that I, trying to match your choice, will go to Wendy's. Hence, I expect you to go to Wendy's, which is why I myself ought to go to Wendy's. Given this distribution of beliefs, it shouldn't matter to me that we've always gone to McDonald's in the past. This doesn't change even if we make explicit the fact that your prediction of what I'm going to do has been based on precedent. Consider the following proposition:

**RB*$_2$* – I believe that you believe that I will go to Wendy's, because I believe that: you know that that's what we've always done, and you believe that people tend to act in the future as they have in the past.

Once I know that precedent-based reasoning has led you to a conclusion about where I'll go, I should not use precedent in predicting your behavior. Once again, my second-order prediction has defeated precedent as a reason for justifying my first-order prediction. The fact that I also happen to know *how* you came to settle on your belief does not change that.

Defeat relations among reasons can be grounded in various ways. For instance, in standard cases, more specific information defeats less specific information (e.g. Horty 2012, 216). To provide just one example, consider the fact that Tweaty is a bird. This is a reason for believing that Tweaty can fly. Suppose further that Tweaty is also a

Penguin which acts as a defeater for the aforementioned inference. Once a reasoner knows that Tweaty is a Penguin she is not justified in concluding that Tweaty flies even though Tweaty is a bird. One might, thus, wonder whether the defeat relation between precedent and higher-order beliefs can be explained in similar ways. I think this is not so. Rather, the defeat relation in our case is simply grounded in basic assumptions about the structure of the game. In a coordination game, as I've explained above, each player is trying to match the other's choice; i.e. each player will act on what she believes the other is going to choose. If the structure of the game, as well as the players' rationality, are common knowledge, then each player knows that the same holds true for the other player. Each player knows that the other will act on her belief about what she thinks the other is going to do, which is why precedent has, at this point, been defeated.

Thus, we can state the intended evidential relations as follows using standard Bayesian notation:

(4) $Pr(RB_n \mid Precedent \wedge RB_{n+1}) = P\,r(RB_n \mid RB_{n+1})$

Condition **(4)** captures the idea that precedent cannot raise the probability of a prediction given the presence of any higher-order belief that would likewise bear on this prediction. The example just stated illustrates that **(4)** is true. But let's expand on this idea a bit more. Consider a case in which I merely believe that you have *some* pertinent reciprocal belief but that I don't know which one. In this case, **(4)** will still be true as the following vignette will make clear:

Fast Food 4. You and I want to meet for lunch. We have two options: McDonald's or Wendy's. We don't care where we'll have lunch as long as we'll have lunch together.

In the past, we've always gone to McDonald's. I learn that a source, who you (perhaps falsely) believe to be infallible, flipped a fair coin, and either told you that I would be at Wendy's this time (if it came up heads), or she told you that I would be at McDonald's (if it came up tails).

In this case, my rational response is debatable. However, one thing is clear; I should not rely on precedent in making my decision. You have a belief about where I'm going to be, and you will try to match my decision based on this belief. This is conclusive for you and, thus, precedent should not be invoked. This is different from the case in which I don't know whether you have any belief about what I'm going to choose. Consider the following case an illustration of this thought:

Fast Food 5. You and I want to meet for lunch. We have two options: McDonald's or Wendy's. We don't care where we'll have lunch as long as we'll have lunch together.

In the past, we've always gone to McDonald's. I learn that a source, who you (perhaps falsely) believe to be infallible, flipped a fair coin, and either told you that I would be at Wendy's this time (if it came up heads), or *she didn't tell you anything at all* (if it came up tails).

In **Fast Food 5**, the intuition that precedent may permissibly be invoked is strong.

The guiding intuition, I think, is that in the case in which she didn't tell you anything at all, you have no belief about what I'm going to do which is why precedent may permissibly be invoked.

These examples show that first-order predictions based on precedent are defeated by the presence of a second-order reciprocal belief. Similar examples can easily be constructed for each level; i.e. *n*th-order predictions based on precedent are defeated by *n+1*th order reciprocal beliefs. The common inductive standard requirement (i.e. expressed in statement **(3)**[39] from above), however, introduces such a higher-order expectation about what the other believes oneself will do, which is why it is implausible.

Now, statement **(3)** frames the common inductive standard in terms of second-order knowledge. However, nothing in the analysis changes if we substitute second-order for *common* knowledge (as Bicchieri has suggested). After all, if the agents have common knowledge about the inductive standard they use, then each level-*n* belief about a player's choice is accompanied by a defeating level *n+1* belief, which is why no level-*n* belief can be justified via precedent.

The cases discussed in this section present the reader with situations in which both types of evidence *conflict*; cases in which precedent favors one response and higher-order beliefs favor a different response. In these cases, precedent is defeated. I should emphasize, however, that conflict is not required for defeat. To see why reconsider **Fast Food 4**; once I learn that you have a belief about where I will be, I simply don't

[39] $K_I(x_y)$ because $K_I(A)$ and $K_I(A \rightsquigarrow x_y)$ and, as one precondition, $K_I(K_y(x_i)$ *because* $Ky(A)$ *and* $K_y(A \rightsquigarrow x_y))$.

care about precedent anymore; precedent becomes a non-issue. This is independent of whether precedent aligns or opposes the conclusions I drew based on higher-order reasoning. For the reader who prefers to reserve "defeat" language for conflict cases, it may be more accurate – although less common – to say that evidence provided by higher-order beliefs "positively screens off" (e.g. Tal and Comesana 2016) precedent-based evidence.

4. Explanatory direction

There is a second objection suggesting that common inductive standards cannot be a precondition for reasoning from precedent in coordination games. Suppose for a minute that two coordinating agents have common inductive standards; i.e. they commonly know that they use precedent as a standard of inference. In this case, we may wonder where this piece of common knowledge came from. We may ask "Why is it that the players have common knowledge that they use *this particular* standard, and not some other, perhaps more outlandish, one?". One compelling candidate answer is that precedent is simply a good standard to use; i.e. it works as a predictor of people's actions. In short, precedent is not a good behavioral predictor because people think that they use it as an inductive standard, but, rather, precedent is a common inductive standard, because it is a good predictor.

Now, if precedent-based inferences were to *presuppose* such common knowledge, then the fact that precedent is simply a good predictor could, on pains of circularity, hardly be recruited for such explanatory purposes. In other words, if precedent were a good standard of inference only if it were already presupposed that the agents

commonly know that this is the standard is invoked, then the fact that that precedent is simply a good predictor could hardly be invoked in explaining why the agents commonly know that this is the standard they invoke. For this reason, predicting behavior using precedent should be allowed without presupposing that the agents commonly know that it is invoked as a standard of inference.

A critic might want to argue that the independent appeal of precedent as a plausible predictor derives from contexts other than coordination games. In coordination games, a player's action, this critic might say, depends on predictions of the other agent's behavior. In other contexts, however, this is not so. For instance, in predicting what wine I'm going to have tonight, you might invoke precedent and predict that I will choose red; and since precedent is a good standard of inference to use in these non-strategic contexts, it is, one might conjecture, also a good standard in the context of coordination games.

I don't think that this thought is plausible though. After all, rational agents should not be inclined to use a standard of inference in a context in which it is simply not suitable. If it were true that reasoning from precedent requires common inductive standards when playing a coordination game, it would simply be a mistake to suppose that such reasoning has independent appeal (i.e. is a plausible standard of inference even without presupposing common inductive standards).

5. *Reasoning from precedent presupposes common belief suspension*

In this chapter, I started with the thought that predicting others' behavior in a coordination game using precedent as an inductive standard seems to be quite

straightforward. This simple inference was captured in statement **(2)**[40]. In the philosophical literature, however, it has been argued that this simple inference ought to be enriched by a common inductive standard; i.e. higher-order expectations about the other players' expectations. The guiding thought throughout this chapter was that this addition is implausible: predicting an agent's behavior in a coordination game using precedent as a standard of inference is incompatible with such higher-order expectations, which is why we should say that precedent-based predictions are permissible only in the absence of higher-order beliefs about the players' choices.

To see how this would go, let's go back to **Fast Food**. When thinking about whether you believe that I will go to McDonald's, I may lack both, the belief that you think that I will, and the belief that I won't go to McDonald's. In short, I might suspend belief about where you think I will go. Belief suspension about a proposition *P* entails, almost everybody agrees, neither believing or disbelieving that *P* is the case (e.g. Bergmann 2005, 420; Wedgwood 2002). Only Jane Friedman has doubts (see Friedman 2017). The connection between suspension and not believing, she contends, is *normative,* not descriptive. An agent who is suspended about *P ought not* believe nor disbelieve it; but since we're operating under the assumption of perfect rationality, we can sidestep these subtleties. Of course, *simply* not believing and disbelieving is not sufficient for belief suspension. After all, a person who has never even considered a certain proposition is not suspended about it; she simply doesn't entertain this proposition. Belief suspension requires some form of cognitive contact

[40] Here a reminder for statement **(2):** $K_I(x_y)$ because $K_I(A)$ and $K_I(A \rightsquigarrow x_y)$.

with the pertinent proposition. It is, as Scott Sturgeon puts it, a state of "committed neutrality" (Sturgeon 2010, 90). The exact form of cognitive contact is, of course, controversial. It has been said that suspending requires "refraining" (Moore 1979), "withholding", or "resisting" believing. This is obviously not the place to adjudicate between these issues; however, I think it is important to keep in mind that an agent may consider a proposition and yet neither believe nor disbelieve it.

Is there a candidate situation in which coordinating agents suspend belief about the other's beliefs concerning one's own choice, suspend belief about what she beliefs oneself beliefs she will do, and so on *ad inf*..? – a situation in which precedent-based predictions go undefeated? I think there is; namely when both agents *commonly know that they have just started deliberating about what to do*.

To start seeing this, we should note that, at the start of their deliberation, the agents haven't considered any evidence yet, which is why they should suspend belief about what the other will do, why she will do it, what the other thinks oneself will do and why she thinks oneself will do it, and so on *ad inf.* In short, at the start of one's deliberation, one has simply not formed any beliefs yet, and one should resist forming these beliefs until one has suitably considered the pertinent evidence. There are two reasons in support of this thought. First, it seems that if attitude suspension is ever appropriate, then this should be when an agent has not considered any evidence. Friedman emphatically states that "it is hard to think of evidential circumstances more appropriate for suspension" than situations in which an agent has no evidence whatsoever. Second, it is reasonable to think that belief suspension is appropriate in

deliberative contexts. This is the position recently defended by Friedman (2017) who states:

> "[W]e can say that there is nothing more to "opening a question in thought" than simply suspending judgment on that question. In suspending about *Q* we make *Q* an object of inquiry. From there we can wonder or be curious or deliberate (and so on) about *Q*. Suspending about a question puts that question on our research agenda." (Friedman 2017, 26)

Deliberating, or "inquiring", about whether *Q* is true is most appropriate when we haven't settled on either believing or disbelieving it. Deliberation is, then, the kind of activity that aims at resolving this neutral state.

Suppose that, in deliberating about how the respective other will act, we're initially suspended, because we haven't considered any evidence, and, as a corollary, have not formed any higher-order belief about the other's choice. Suppose next, that we (commonly) know that we've just started our deliberative process, and, thus, (commonly) know that we are so-suspended. In this case, we don't have any higher-order beliefs about where the other thinks oneself will go. I don't have any belief about what you think I will do, and I also know that you don't have any such belief about what I think you will do. In this initial state, there are no higher-order beliefs. All potentially defeating higher-order beliefs are absent and we may, at least as far as the relevant defeaters go, permissibly predict the respective other's behavior using precedent as a standard of inference.

The following picture emerges: It is epistemically permissible for agents to predict each other's behavior based on precedent only in the absence of higher-order reciprocal beliefs about their actions. The latter condition is (for instance) satisfied when agents commonly know that they've just started deliberating about how to act and are thus suspended about what the other thinks oneself will do, what she thinks oneself thinks the other will do etc.

6. *Conclusion*

David Lewis wrote that salience (e.g. grounded in precedent) can support coordination by providing reasons for choosing a strategy when there is "no stronger ground for choice" (Lewis 1969, 35). Higher-order predictions about what the other player thinks oneself will choose present, I have argued in this chapter, such a "stronger ground for choice". For this reason, precedent-based predictions are legitimate only in the *absence of such higher-order behavioral predictions.* More concretely, I have argued that this absence requirement is satisfied when the agents commonly know that they both suspend belief about what the respective other is going to do and why she's going to do it. This claim is directed against a philosophical doctrine according to which precedent-based predictions require a "common inductive standard"; e.g. higher-order predictions about what the other player thinks oneself will choose.

The idea that common knowledge requirements should, in the context of coordination games, by couched in terms of belief *absences* has rarely been noticed. In fact, I only know of two authors who have recognized such absences to be relevant

in spelling out common knowledge in coordination games. First, in a discussion "mutual" knowledge in communicative contexts, Martin Davies (1987, 717) suggests that "the philosophical work which was to be done by the notion of mutual knowledge should instead be assigned to a negatively characterized notion: mutual absence of doubt." The second reference is from Richard Moore's (2013, 492) discussion of common knowledge in conventional behavior. He notes that "the extent to which common knowledge is necessary for conventional activity will be determined by its coordinative role. Such a role might consist in *protecting participants in a convention from higher-order doubts* about the conformity of others" (my italics). These somewhat cursory remarks merely hint at the structural importance belief-absences have for solving coordination games. In this chapter, I've tried to elaborate on this idea. Importantly, the present analysis showed that common knowledge was not simply an unnecessarily baroque theoretical element, but, rather it's presence was said to act as a defeater for reasoning from precedent in the context of solving coordination games.

Appendix A: Beyond 'Interaction': How to Understand Social Effects on Social Cognition

1. *Introduction*

Consider the following example of a typical social interaction:

> **Ian and Mia.** Mia enters a coffee shop and sees her best friend Ian sitting on the sofa. Ian doesn't notice her right away because he is stooped over his phone, closely examining the image of a woman on a dating website. Ian looks up and sees Mia, who smirks when she sees what he's been looking at. Ian blushes and quickly puts away his phone. 'It's not what you think', he says. 'I'm helping Sarah set up her profile'. Mia chuckles and asks Ian if he would like something from the barista. Ian asks for a green tea. On her way back from the counter, Mia trips, and spills both of their drinks all over her jeans. She looks around, and notices how everybody in the coffee shop is staring at her.

Now, contrast this with the following description of a standard false-belief task procedure, which is typical of social cognition research:

> **Standard False-Belief Task.** Children see a toy figure of a boy and a sheet of paper with a backpack and a closet drawn on it. 'Here's Scott. Scott wants to find his mittens. His mittens might be in his backpack or they might be in the closet. Really, Scott's mittens are in his backpack. But Scott thinks his mittens are in the closet'. 'So, where will Scott look for his mittens? In his backpack or in the closet?' (the target question) 'Where are Scott's

> mittens really? In his backpack or in the closet?' (the reality question). To be correct the child must answer the target question 'closet' and answer the reality question 'backpack'. (Wellman and Liu 2004)

Real-life social interactions like **Ian and Mia** are complex. They involve, among other things, belief ascriptions, gaze cues, emotional signals, gestures, relationships, and social conventions. Despite this complexity, the scientific study of such situations tends to rely on simplified, highly artificial paradigms like **Standard False-Belief Task**. Ostensibly, the kind of knowledge being tested in the false-belief task is also supposed to be the knowledge that Ian and Mia use in order to successfully navigate their social encounter—namely, their theory-of-mind. However, the difference between these two vignettes is hard to ignore. Of course, some might argue that for all their artificiality, we need tools like the false-belief task if we are ever to begin to make sense of how social cognition functions. This is a trade-off inherent to all experimental psychology: if we desire scientific rigor, we must sacrifice some ecological validity.

However, there are a number of theorists who think that experimental paradigms in social cognition research like **Standard False-Belief Task** sacrifice far too much (De Jaegher and Di Paolo 2007; Gallagher and Hutto 2008; Schilbach *et al.* 2013). For instance, in this experiment the child is set apart from Scott. There is no possibility for the two to interact. There are no reciprocal gaze cues, no emotional signals, no gestures, and no relationships. The child is merely a passive observer. In **Ian and Mia**, on the other hand, both agents are interacting. They mutually respond to and

transmit a wide range of social cues, which get interpreted in a context-sensitive fashion.

Interactionists conclude that, as a consequence of experimental oversimplification, traditional research on human social cognition has lost sight of the very phenomenon it set out to explain. What is needed, they propose, is an 'interactive turn' towards more 'second-personal' methods and theories that acknowledge the dynamic, interdependent aspects of ordinary social experiences. More specifically, according to interactionists, past research is problematic because it relies heavily upon observational experimental paradigms. Real social cognition, however, almost always takes place in interactive contexts. As a result, current theoretical and empirical paradigms are thought to be ill-suited to study the cognitive processes at work in real-life social activities.

One example of this kind of 'interactionist' approach to social cognition research is the double TV monitor paradigm (Murray and Trevarthen 1985). In this experiment, 2-month old infants were shown a TV screen displaying a video of their mothers. In the 'interactive' condition, the video was live, while in the 'non-interactive' condition, the video showed a replay of their mother's actions. It was found that infants quickly disengaged when presented with the replay video, but were far more motivated to attend to the feed in the interactive condition.

Another example is the perceptual crossing paradigm (Auvray *et al.* 2009; Auvray and Rohde 2012). In this experiment, two players move an avatar along a one-dimensional strip using a computer mouse. When moving her own avatar along the

strip, a player can cross paths with three objects: a static object, the other player's avatar, and the other player's avatar's shadow (i.e. an object copying the movements of the other player's avatar). Each agent receives the same sensory feedback upon crossing paths with any of these three objects. Importantly, when one player's avatar crosses paths with another player's shadow only the player with the avatar receives feedback. If two players' avatars meet, both players receive sensory feedback. Interestingly, although the sensory feedback a player receives from crossing paths with any of the objects is identical, players nevertheless typically manage to 'find' one another (i.e. oscillate their avatars around each other).

The interactionist criticism, thus far, amounts to the claim that traditional research paradigms such as **Standard False-Belief Task** need to be supplemented by novel interactive paradigms; and this criticism is well-taken. However, some interactionists have taken their critique a step further, and argued that the socio-cognitive processes at work in interactive contexts are fundamentally distinct from those that operate in observational ones. For instance, Shaun Gallagher and Daniel Hutto have argued at length that mental state attribution really only occurs in observational scenarios; in social interactions, we rely upon a range of non-mentalistic processes, including gaze-following, social narratives, and emotional mirroring (Gallagher 2001, 2009; Hutto 2004, 2007; Gallagher and Povinelli 2012).

At times, interactionists in the enactivist tradition make the even more radical claim that social interactions can actually constitute social cognition. Interactions, it is argued, have emergent properties that cannot be reduced to contributions of

individuals. When two autonomous agents act in such a manner that their actions are 'coupled' (i.e. causally interdependent), this can create a higher order 'dynamical system' with its own intrinsic properties. These systems, it is claimed, are the true loci of social cognition (De Jaegher *et al.* 2010). We should, therefore, abandon the idea that social cognition can fully be explained in purely individualistic terms. Instead, according to the interactionists, social cognition researchers ought to focus their efforts on the intrinsic properties of interactive systems.

Other proponents of the interactive turn have de-emphasized the claim that social interactions are constitutive of social cognition. For instance, Gallotti and Frith have argued that interacting agents have 'novel routes to knowledge of other minds' that facilitate cooperation and team reasoning (Gallotti and Frith 2013, 162). This route to social knowledge is achieved by entering into the 'we-mode', a psychological state in which aspects of an interactive scene are represented via distinctively collective mental attitudes: believing-together, intending-together, desiring-together, etc. When agents enter the 'we-mode', they co-represent the action-possibilities available to their interactive partners, and use this information to make decisions that achieve collective ends. Andreas Roepstorff and colleagues have also proposed that social situations can be interactive to varying degrees; with increasing degrees of interactivity, they find corresponding effects upon processing speed (Tylén *et al.* 2012), accurate collective decision-making (Bahrami *et al.* 2012), and physiological and behavioural alignment (Fusaroli *et al.* 2016).

While proponents of the interactive turn come in various flavors, they all endorse a central methodological claim: in order to promote ecological validity, experiments in social cognition need to become less observational and more interactive. In this chapter, we will argue that this way of thinking is misguided. We are of course in favor of improving the ecological validity of social cognition research; however, we think that the notion of 'social interaction', as it is currently being deployed, is the wrong tool for the job. We argue that contrasting social cognition in interactive and non-interactive contexts is often uninformative, and prone to methodological confusion. This is because both the proximal causes and underlying mechanisms that support naturalistic social cognition tend to straddle the interaction/observation dichotomy. In short, we believe that emphasizing 'interaction' is a red herring.

To show why this is the case, we will first turn our attention to the definition of 'interaction' that has become the standard in the interactionist literature. We will argue that this definition introduces concepts that needlessly complicate the target phenomena. In its place, we will offer a pared down, minimalist definition of 'interaction' that adequately captures the phenomena that interactionists are interested in.

Next, we will point out an obstacle to any cognitive scientist wishing to implement 'interactionist' experimental paradigms. This is that interactions are typically composed of many different social elements that are not themselves interactive. These concomitant social elements create a number of potential confounds for interactionist experiments, which social cognition researchers would do well to control for. To this

end, we review four bodies of literature that illustrate the need for appropriate, non-interactive controls in interactionist paradigms: the 'Social Simon Effect', spontaneous perspective-taking, imitation, and conversational alignment.

Finally, we will argue that in many cases, so-called 'interactionist' paradigms have really featured ersatz interactions. We think this shows that it is not interaction as such that really makes a difference in social cognition research, but rather that individual participants believe themselves to be interacting. This contradicts the basic anti-individualist thrust of interactionism.

2. Defining "interaction"

We now turn to the issue of defining 'social interaction'. This turns out to be a delicate matter: while it is widely acknowledged that to develop an adequate theory of social cognition, we should be studying social interactions, there are ways of defining the term that largely presuppose a particular theory of social cognition. But if studying social interaction is supposed to provide evidence for these same theories, this ends up being circular. What is needed, rather, is a theory-neutral definition of social interaction that all interested parties can agree upon. This notion of interaction can then serve as a common point of departure for future debates. Therefore, our strategy in this section will be to start with the most prominent definition of social interaction in the extant literature, and then pare it down to a minimal, theory-neutral form.

The most influential definition of 'social interaction' comes from De Jaegher, Di Paolo, and Gallagher 2010:

> **De Jaegher Interaction.** Two or more autonomous agents co-regulating their coupling with the effect that their autonomy is not destroyed and their relational dynamics acquire an autonomy of their own. Examples: conversations, collaborative work, arguments, collective action, dancing and so on. (De Jaegher *et al.* 2010, 441)

An 'autonomous system' is further defined as a 'network of co-dependent, precarious processes able to sustain itself and define an identity as a self-determined system' (De Jaegher *et al.* 2010, 441). The set of autonomous systems, on this definition, includes most biological life-forms, from single-celled organisms to human beings, and also socially constructed entities, like corporations and nations. In the context of social cognition, the relevant class of autonomous systems is restricted to autonomous agents. 'Coupling' occurs when one autonomous system causally impacts the functioning of another. Coupling is said to be 'regulated' when this causal impact is in some way controlled by that system; and it is said to be 'co-regulated' when two or more autonomous systems are controlling how they causally impact one another. Genuine social interactions, on this view, occur when this co-regulated coupling results in the creation of a new autonomous system while still preserving the autonomy of the co-regulators. Lastly, this emerging interactive system is required to be temporally extended enough to take on 'autonomy' of its own.

Our first issue with this definition is related to the idea that genuine social interactions take on 'an autonomy their own'. As noted above, a definition of 'social interaction' should, where possible, be theory-neutral; it should not entail a particular

social ontology. However, the ontology implied by the above phrase is highly controversial: namely, that interactions create new autonomous systems. These autonomous systems are then thought to form the proper objects of social cognition research: they literally constitute social cognition (De Jaegher *et al.* 2010). But a number of authors have argued that this claim amounts to a confusion of constitution and causation (Herschbach 2012; Carruthers 2015). Given that this debate is still ongoing, it seems unnecessary to hardwire such a controversial metaphysical claim into a practical, theory-neutral definition. Therefore, we propose the following first initial revision to De Jaegher's *et al.*'s definition:

> **Interaction - First Revision.** Two or more autonomous agents co-regulating their coupling with the effect that their autonomy is not destroyed.

Our second worry concerns the role that the concept of 'autonomy' plays in this definition. In De Jaegher and colleagues' definition, 'autonomy' is introduced as a technical notion according to which almost all biological life forms, not just human beings, can constitute autonomous systems (i.e. they can form self-sustaining and self-determining systems). Likewise, interactions between such autonomous systems don't necessarily have to involve human beings either: interactive systems would come into being whenever two cells cross paths in a petri dish, and whenever two countries engage in diplomatic negotiations. With such a broad scope, one might worry that this notion of social interaction is indeed too broad to be of any scientific utility. If the study of social cognition is to take an 'interactive turn', then interaction

needs to be something that can be operationalized in a controlled, experimental setting.

Presumably, it is for these reasons that De Jaegher and colleagues narrow their definition to be specifically about autonomous 'agents'. However, in this case 'autonomy'—at least in the technical sense of the term—does not do any definitional work. This is because the set of agents is a proper subset of the set of autonomous systems. Therefore, the phrase 'autonomous agent' is not more informative than the term 'agent'.

Furthermore, given their technical notion of autonomy, it is unclear why cases of coercion should be discounted, as De Jaegher and colleagues maintain (De Jaegher and Di Paolo 2007, 495; De Jaegher *et al*. 2010, 443). In a case of armed robbery, for instance, it would seem that we have an instance of correlated mutual behaviour that is at least as complex as the case of two people having a conversation. Why, then, would this fail to create an interaction? According to De Jaegher and colleagues, the coercive nature of the mugger's actions would 'destroy the autonomy' of the victim. If the criteria for autonomy are so weak that bacteria in a petri dish can form an autonomous system, it is hard to see how it could be destroyed simply by demanding, 'Your money or your life!'. Even if the victim complies, it seems as though her status as an autonomous system in the sense being used here would be preserved.

Of course, there is a classic, Kantian sense in which the victim's autonomy in this situation is compromised—namely, her ability to act in accordance with a law of her own choosing. If interactionists were to adopt this notion of autonomy in their

definition, they could avoid the charge of vacuity. However, we would then need to dramatically revise the range of cases that would count as social interactions. First, the subset of entities that possess autonomy in this strong sense will be much smaller than those that possess it in the weaker sense. Young children and animals, for instance, are unlikely to be autonomous in this sense. Drug addicts and persons with cognitive disabilities would also likely to fall below the threshold. Women in highly patriarchal countries with oppressive religious laws would also lack this kind of autonomy. Second, although human agents can be autonomous in this sense, it is unclear what it would mean for a co-regulated coupling to create an autonomous system. In short, it is not clear when—if ever—the conditions for interaction would obtain, given this notion of autonomy. Lastly, and most importantly, it is not at all clear why an obviously normative notion should play a role in cognitive science. The fact that a person cannot act in accordance with the law of her own choosing does not obviously bear on the cognitive mechanisms she brings to bear when encountering other agents.

These problems associated with the Kantian notion of autonomy also generalize to other normative theories of autonomy, which are generally unfit to constrain cognitive theories of interaction. To see this more clearly, consider the higher-order theory of autonomy defended by Michael Bratman (Bratman 2003, 2007). According to Bratman, autonomous agents treat mere considerations to act as justifying reasons to act (2007, 178). Treating one's considerations in this way functions as a guide to resolve indecision and is, therefore, desirable. Autonomy, understood in this way, is a normative notion. Agents can fail to act autonomously, if they fail to have appropriate

higher-order regard for their first-order motivations. Importantly, it is implausible that agents who fail to treat their considerations for action as justifying reasons cannot engage in mundane (but clear) forms of interaction (e.g. paying the cashier for the groceries I wish to buy). In short, normative theories of autonomy introduce constraints that are too restrictive to ground cognitive accounts of interaction.

Thus, De Jaegher and colleagues' reliance on 'autonomy' in their definition faces a dilemma: given the original, more technical notion of autonomy, interactions are so ubiquitous and variable that they do not form a category of scientific interest. Given a more demanding, normative notion of autonomy, interactions become so rare that it is not clear whether they occur at all. Interactionists could address this issue by providing an alternative account of 'autonomous systems' that is situated somewhere in between these two extremes. But until such an account is provided, the notion of 'autonomy' is not scientifically useful. Therefore, we propose a second revision to De Jaegher's definition:

> **Interaction - Second Revision.** Co-regulated coupling between conscious human beings.

This revised definition does away with the notion that interactions must be performed by autonomous systems. But nothing serious is lost. We noted that once the relevant class of agents is specified, the further classification 'autonomous agents' is explanatorily inert. The revised definition makes explicit that, in the context of social cognition, the relevant class of agents are conscious human beings. To be sure, other types of organism may also engage in interactions, but this need not concern us.

Lastly, we propose a small addition to our definition: two agents or more co-regulate their coupling if the actors knowingly[41] affect each other's actions. This further specification is necessary to rule out cases in which agents affect each other's actions by mere accident. Consider the case in which you swipe the foliage from your lawn into my lawn. I, thinking that a sudden gust of wind is responsible, swipe it back into your lawn. You, having the same thought, swipe it back into my lawn. We keep doing this until the end of August, when the foliage finally decays. Although we're affecting each other, we are, intuitively, not interacting. Moreover, our behaviour is uninteresting from the perspective of social psychology. Lastly, the addition of knowingly is preferable to the addition intentionally, because it does not exclude cases in which several agents affect each other's actions by mere foresight[42].

In summary: after a few clarificatory modifications of De Jaegher and colleagues' account of interaction, we are left with the following definition.

> **Minimal Social Interaction.** When two or more conscious human beings mutually and knowingly affect one another's actions, they are engaged in a social interaction.

This minimalist definition fits nicely with paradigmatic examples of social interaction: conversation, dancing, cooking a meal together, playing tennis, etc. It also

[41] Note that, for our purposes, knowingly should be given a deflationary reading that is common in psychology (Dienes and Perner 1999; Nagel 2013). Knowing X, in this sense, means 'being aware of X and being sensitive to X when acting'. For instance, for Dienes and Perner (1999) mere perceptual awareness is sufficient for knowledge. What is more, having knowledge does not require recognizing that one has knowledge; i.e. it does not presuppose the concept KNOWLEDGE. Lastly, interacting knowingly does not presuppose the concept of INTERACTION; rather, it merely requires being aware of the constituents of interaction (e.g. the other person's voice and actions).

[42] Think, for instance, of a case in which you merely intend to get the foliage off your lawn, but you also foresee that I'll be mad when I find the foliage on my lawn. However, you don't intend to make me mad; you merely foresee that this will happen.

does not, however, eliminate cases of coercion and manipulation, such as the mugger scenario, or even actively violent encounters, such as fistfights. But it is not clear why these cases should be eliminated: surely, not all social interactions are pleasant and cooperative. While we may morally disapprove of these actions, this does not make them any less interactive.

This minimalist definition also fits nicely with key examples of interactionist experiments. In the Double TV-monitor paradigm, for instance, the live-feed condition makes it so that infants and their mothers are able to mutually respond to one another's actions; when the recording of the mother's expressions are played back for the child, this is no longer possible. In the perceptual-crossing study, participants are able to locate one another's sensors on the one-dimensional strip because they are able to mutually respond to one another, whereas the 'shadow' and the fixed object cannot.

According to the minimal approach, paradigms like the standard false-belief task would not count as interactive. This is because the actions of the character in the vignette do not affect the child's actions, and the child's actions do not affect those of the character in the vignette. The child merely observes the events taking place in the vignette, and then makes a prediction about them. There is no opportunity for a reciprocal exchange of information between the child and the character, nor any possibility for mutuality. It is decidedly non-interactive.

With this definition in hand, we are now in a position to defend our main point: if we want to improve the ecological validity of social cognition research, we should not

frame this effort in terms of a distinction between interactive and observational scenarios.

3. The constituents of interaction

Proponents of an 'interactive turn' in social cognition research claim that in order to learn more about the nature of social cognition, we need to create more interactive experimental designs, and get away from purely observational paradigms. There is nothing wrong with designing interactive paradigms; however, it's not clear how much we really learn when we try to directly compare interactive and non-interactive contexts. This is because social interactions typically involve many different elements that are not themselves interactive.

To illustrate, take a prototypical interaction: a conversation with a colleague by the drinking fountain. Such an encounter would involve the physical co-presence of two individuals; however, this by itself would not make it an interaction. Likewise, the two speakers might possess mutual background knowledge about one another, including beliefs about each other's occupation, political views, short- and long-term goals, and so on. But this too does not make the encounter an interactive one. The conversation also involves the use of language. But even this, all by itself, fails to make the context interactive: one could easily imagine a person speaking aloud to herself, while another person ignores her. None of these elements, by themselves, it seems, are enough to make an encounter interactive. But all the same, they seem to be very important elements of the context, from a cognitive perspective.

Social interactions like this one seem to be complex events, composed of many elements that contribute to its interactive nature, and yet are not themselves interactive. All of these elements—physical co-presence, background knowledge, the use of language—often co-occur in social interactions, but are neither necessary nor sufficient for an interaction to occur. But, as we shall see in this section, they still have considerable effects on social cognition. As such, it is unclear whether 'interactive' effects on social cognition are driven by interaction as such, or by one of its component elements. In this section, we use several distinct bodies of evidence to argue that simply contrasting interactive and non-interactive scenarios is not informative. This, we claim, reveals a key oversight in the interactionist approach.

3.1. The Social Simon Effect (Sebanz *et al*. 2003)

In a typical 'Simon' task, subjects carry out responses using their left and right hands to stimuli appearing on the left and right sides of a screen; typically, subjects are faster to respond to stimuli appearing on the side congruent with the response (i.e. left side of the screen with left hand response), and slower to respond to items appearing on the incongruent side (i.e. left side of the screen with right hand response) (Craft and Simon 1970). Natalie Sebanz and colleagues modified this task so that it involved two subjects participating in parallel to one another, each responsible for responding with either the left or right hand; thus, subjects only had to respond in a Go/No-Go fashion depending on what they saw on the screen, regardless of which side the stimuli appeared on (Sebanz *et al*. 2003). Importantly, their performance in no way depended upon what the other agent did—all they ever had to do was pay attention to

their own screen and respond accordingly. Thus, there was nothing interactive about the task.

When subjects performed this task alone in a control condition, there was no spatial congruency effect—they were equally quick to respond to items on either side of the screen. But in the social condition, there was a spatial congruency effect: subjects were slower to respond to items on the side opposite their response hand (and on the same side as the other participant's response hand). In effect, the presence of another agent altered the way they represented their environment, such that it included both their own action affordances, and those of the other agent. Even when seated side-by-side with another agent completing totally independent tasks, their sheer presence affects how we represent and respond to the environment.

Since Sebanz and colleagues discovered the Social Simon Effect, a number of other experiments using similar paradigms have replicated and extended this finding. Using variants of the Social Simon paradigm, Guagnano and colleagues found that the Social Simon Effect dissipated with increased spatial separation between the two agents (i.e. within or beyond arm's length) (Guagnano *et al.* 2010); Vlainic and colleagues found that the effect persisted even when subjects had no online perceptual feedback from the other participant, demonstrating that simply knowing that another agent is completing a similar task is enough to alter how one represents one's own action space (Vlainic *et al.* 2010). Freundlieb and colleagues showed that when

another agent was co-present but inactive, or co-present but completing a task of which the subject was ignorant, the effect dissipates (Freundlieb *et al.* 2015).[43]

Thus, simply knowing that another agent is acting nearby is enough to alter the way that we respond to our environment, even when no interaction—even in the minimal sense—is taking place. Given that most interactive experimental designs include the co-presence of active agents, it may be that co-presence effects—which are not, in fact, the products of interaction—also occur in those tasks. This creates a methodological confound for proponents of the 'interactive turn' in experimental design: how are we to know whether purported interaction effects are genuine, or simply the product of the co-presence of other active agents?

3.2. Level-2 perspective taking

Physical co-presence also seems to have an effect upon whether or not we spontaneously engage in certain forms of perspective-taking, the representation of what another agent can see. Psychologists typically distinguish between two 'levels' of perspective-taking (Masangkay *et al.* 1974; Flavell *et al.* 1981): Level-1 perspective-taking means representing whether or not a particular object is in the visual field of an agent, and is sensitive to external, environmental factors like line-of-sight and occlusion (Michelon and Zacks 2006). Level-2 perspective-taking further involves the ability to represent how an object appears to another agent; for instance, the numeral '6' might, from one angle, appear to represent the number six, and from

[43] Guagnano *et al*. (2010) interpret their results as showing that the Social Simon Effect is due to participants representing their own action space, not the action affordances of those around them. But this claim in undermined by the results of Vlainic *et al.* (2010) and Freundlieb *et al*. (2015), which show that knowledge of another agent's action is key to generating the spatial congruency effect.

another angle, appear to represent the number nine; sensitivity to these differences requires an understanding of the aspectual nature of perception (Surtees *et al*. 2012, 2016). Until recently, our best evidence suggested that while Level-1 perspective-taking is automatic and effortless, Level-2 perspective-taking is effortful and requires top-down, intentional control (Qureshi *et al.* 2010; Samson *et al.* 2010; Surtees *et al.* 2012). However, the relevant perspective being taken in these tasks was always that of a non-descript, computer generated avatar. But when the avatar is replaced with a live agent, we see a very different effect (Elekes *et al.* 2016).

In this experiment, subjects sat in front of a monitor lying flat in front of them, and had to verify whether or not the numeral on the screen matched a number they heard in an audio recording. In the Individual condition, subjects completed this task alone; in the Joint condition, subjects sat opposite another participant who was either also completing a number-verification task (i.e. the perspective-dependent task), or a different task in which they had to say whether the colour of the numeral on the screen was the same as one they'd seen just previously (i.e. the perspective-independent task). Participants in the joint condition always knew which task the person opposite them was completing. Importantly, all subjects had to do was complete their own task—the actions of the other agent were always irrelevant. Thus, the task was not interactive (given our definition).

Elekes and colleagues found that subjects in the Joint condition were slower and made more errors than in the Individual condition, but only when both completed the perspective-dependent task and the numerals of the screen were such that their values

differed on the basis of perspective (i.e. 2, 5, 6 and 9); for numerals whose values appeared to be the same regardless of which side of the table the participant was at (i.e. 0 and 8), there was no difference between the Individual and Joint conditions. In effect, subjects were only slower when 1) they had a live partner, 2) they believed that their partner had a similar goal, and 3) the partner's response would diverge from their own on the basis of their Level-2 perspective. In other words, when subjects knew that the person across the table from them was viewing the numeral on the screen as a number, they spontaneously maintained a representation of what he or she saw, and this representation then interfered with their own performance.

Thus, in this task, the mere co-presence of an active agent was not sufficient to prompt Level-2 perspective-taking, but the combination of co-presence and the knowledge that this agent had a goal similar to their own did. These results complement those of the Social Simon Task: when another agent is co-present, active, and has a goal similar to our own, we spontaneously represent both how the environment appears to them, and the kinds of actions that are available to them in that environment.

In interactive scenarios, of course, we are usually aware of the physical presence of other agents and their goals. Thus, we might expect that in those scenarios, we would also represent the affordances of the environment differently, or spontaneously adopt our partner's visual perspective. Upon observing all of these levels of socio-cognitive processing layered on top of one another, it is tempting to hypothesize that social interactions are irreducibly complex, and possess emergent properties.

However, many of the constituents of this interaction are indeed isolable, and we can study the effects of these constituents individually. Moreover, we know that these social effects on social cognition are not inherently interactive, because we can also observe them in non-interactive scenarios. This is, we think, the central problem with the 'interactive turn': by focusing on interaction as a global property of social-cognitive scenarios, we miss out on a wealth of local, fine-grained information that may be present in non-interactive contexts.

A proponent of the 'interactive turn' could object that the cases we've described here are in fact best understood as effects of 'we-mode' cognition (Gallotti and Frith 2013). For even though subjects in the Social Simon tasks and the perspective-taking task are not yet engaging in an interaction, they may be cognitively preparing for an interaction. The sheer proximity of their partners and the similarity of their tasks, the interactionist might argue, creates the sense that they are about to interact with one another, and this leads them to become more sensitive to their partner's perspective and action possibilities. Alternatively, these contexts might be said to create the illusion of interaction, where in fact there is none. Either way, the objection might go, these effects really only make sense in an interactionist framework.

We think that this objection makes an important point, but also a crucial concession. It may well be true that the cognitive processing that takes place in these near-interactive contexts have the function of supporting interaction. However, the fact remains that their presence was revealed in a non-interactive context, and that interaction is not necessary for eliciting them. Rather, the non-interactive task-design

was a crucial part of discovering these processes. Thus, even if interaction might be a part of the explanation of why these effects are present, it was crucial that interaction was not a part of the task that revealed them.

In sum, it is important to identify the various sub-components of interaction, and not to mistake the effects of these sub-components for effects of the interaction itself. In practice, this will mean employing experimental paradigms that are explicitly non-interactive.

3.3. Interaction effects on infant learning

One line of research that seems to emphasize the importance of interactive methods is the literature on 'natural pedagogy' (Gergely and Csibra 2005; Csibra and Gergely 2006, 2009b). According to this view, when an infant is addressed with certain ostensive signals (e.g. eyebrow-raising, eye contact, infant-directed speech), children spontaneously adopt a specialized learning stance. This learning stance prepares children to attend to certain kinds of information, such as facts about the identity and category-membership (Csibra and Gergely 2009a). The pedagogical stance is also said to facilitate imitative learning.

The natural pedagogy hypothesis is not an explicitly interactionist proposal. However, it does seem to buy into the central methodological prescription of interactionism: there are certain forms of cognition that can only be studied in interactive contexts. Experiments in this tradition also frequently use observational controls to demonstrate the effects of pedagogical learning. For example, Yoon and colleagues found that 9-month olds tended to encode information about the location

of an object in a non-interactive context, but instead encoded information about the object's identity in an interactive context with pedagogical cues (i.e. where an experimenter engaged in infant-directed speech and eye-contact) *(*Yoon *et al.* 2008). The authors suggest that this is because interactive, pedagogical contexts prompt children to pay special attention to generic information. Likewise, in a study with 14- to 16-month-olds, Brugger and colleagues found that infants were more likely to imitate novel actions more in interactive, pedagogical contexts than in observational contexts (Brugger *et al.* 2007). Based on these contrastive observational-versus-interactive designs, proponents of the natural pedagogy hypothesis argue that pedagogical interactions trigger specialized learning mechanisms that are not active in observational contexts.

The natural pedagogy hypothesis, however, remains controversial if cast as a theory specifically about interaction. To see why, note that in Brugger *et al*. (2007) and Yoon *et al*. (2008), the non-interactive condition was both non-communicative (i.e. the action was demonstrated by a solitary person) and observational (i.e. the child was not addressed through ostensive cues). Communicative contexts, however, are not necessarily interactive: one can observe communication between third parties without actively participating in it. Hence, these experiments leave open the possibility that the same learning effects attributed to pedagogical interactions might also occur in observational but communicative contexts.

Once the relevant distinctions are introduced, the importance of interaction in imitative learning becomes much less obvious. For instance, (Matheson *et al*. 2013)

conducted a study in which 18-month-olds and 24-month-olds imitated novel actions (e.g. ringing a doorbell using one's forehead) in (a) an interactive condition in which the experimenter addressed the infant using typical ostensive cues, (b) an observational and non-communicative condition in which the infant watched the experimenter perform the novel action all by herself, and (c) an observational-communicative condition, in which the infant watched the experimenter perform the novel actions while demonstrating them to another person. They found that 18-month-olds imitated more in the interactive condition than in the observational–non-communicative condition, but not significantly more than in the observational–communicative condition. In other words, it was the communicative dimension of the interactive condition that seemed to have improved imitation, rather than interaction as such. In 24-month-olds, meanwhile, there were no differences across all three conditions.[44]

Shimpi and colleagues achieved a similar result while also manipulating the child's familiarity with the imitative model (e.g. whether the model was a family member, a complete stranger, or a stranger with whom the child had briefly interacted[45] before the task began) (Shimpi *et al.* 2013). Interestingly, children in the observational-communicative condition imitated consistently regardless of whether they were familiar with the model; in contrast, children in the interactive condition imitated far less with unfamiliar models than familiar models. Thus, while children were quite

[44] Interestingly, emulation was significantly higher in the solitary–non-communicative condition than in the interactive condition. (An actor's action is said to be emulated by an agent, if the actor's goal is copied by the action. An action is said to be imitated, if the agent copies the actor's exact action sequence.)
[45] Familiarity was established through a 10-minute warm-up period in which the experimenter played a sorting game with the child.

adept at learning imitate complete strangers in observational–communicative contexts, some familiarity with the model was a prerequisite for imitative learning in interactive contexts.

On the one hand, these experiments do suggest that interaction can facilitate imitative learning in infants. However, these effects are not particularly pronounced: in Matheson *et al*. (2013) imitative learning in the older children was the same for all three conditions; in Shimpi *et al*. (2013) observational learning in communicative contexts was robust; and interactive learning was crucially dependent on the familiarity of the actor. The importance of interaction in imitative learning thus appears to be overstated. Similar observational-communicative controls have yet to be carried out for other forms of learning described by the natural pedagogy hypothesis (e.g. generic learning), and we cannot say for certain whether observational learning will be equally robust in that domain. However, we think there is good reason to find out.

3.4. Conversational alignment

We've noted that there are several social factors that are present in many social interactions, that have noticeable effects on social cognition, and that might be mistaken for interaction effects, but which are in fact non-interactive. However, an interactionist might object, even if these factors are present in non-interactive scenarios, they may still have unique effects in the context of a social interaction. Take, for instance, our paradigmatic example of a social interaction: conversation. We have pointed out that language use, by itself, is not inherently interactive. But, the

interactionist might insist, language works much differently when studied as monologue than when it is studied as dialogue.

This is the central point behind the 'interactive alignment' research program of Martin Pickering and Simon Garrod, which has focused on the nature of speech production and comprehension during naturalistic dialogues (Garrod and Pickering 2004, 2009; Pickering and Garrod 2004, 2013). Explicit in this research program is a critique of psycholinguistic theories based on the study of comprehension and production of speech in non-interactive contexts (i.e. monologue). The most natural and basic form of language use, they claim, is dialogue; to develop a full understanding of the mechanisms of language, we need to study it in this form.

Central to Pickering and Garrod's positive account is the observation that speakers in a dialogue will tend to converge upon matching representations at the lexical, semantic, and syntactic levels—a phenomenon the authors call 'conversational alignment'. For instance, syntactic alignment refers to the spontaneous tendency of a speaker to use a particular syntactic construction when that same construction has just been used by an interlocutor (e.g. the cowboy gives the pirate a banana versus the pirate gives the banana to the pirate (Pickering and Branigan 1999; Branigan *et al.* 2000, 2007). In dialogue, this alignment of representations is said to take place at multiple levels simultaneously, with alignment at one level facilitating alignment at other levels through the co-activation of multi-level associative networks. As a result of this alignment process, participants in a dialogue achieve a high level of communicative fluency. This enables them to rapidly recover meaning from each

other's utterances, even when these utterances are otherwise fragmentary, overlapping, and entirely ungrammatical. Other researchers have also extended the study of alignment in dialogue beyond the coordination of linguistic representations, and found evidence for analogous forms of synchronization in eye movements (Dale *et al.* 2011) and heart-rate (Fusaroli *et al.* 2016).

We agree with the general project of studying dialogue in naturalistic circumstances. However, we argue that much of Pickering and Garrod's own account of the mechanisms supporting conversational alignment depends upon evidence from individualistic paradigms. Moreover, while there are some differences in the magnitude of the relevant effects when these are measured in interactive contexts, these differences are readily explained in terms of other non-interactive mechanisms, such as increased attention. Finally, even where we do find uniquely interactive alignment effects, individualistic mechanisms still play an important role in their explanation.

For instance, Garrod and Pickering have suggested that alignment between speakers and listeners is a product of representational processes that are shared between the comprehension and production systems. Thus, when a listener hears an utterance of a sentence with a certain syntactic form or lexical item, those representations are primed for use in speech production. However, much of the evidence that Pickering and Garrod present for this mechanistic hypothesis is derived from non-interactive tasks (i.e. 'monologue'). For instance, the 'structural persistence' or priming of syntactic forms from comprehension to production has

been established in numerous individualistic experimental paradigms, which Pickering and Garrod cite as evidence (Bock 1986; Bock *et al.* 2007). Pickering and Garrod (2013) also suggest that the shared representational processes in comprehension and production are the product of forward modelling mechanisms for action-planning (Davidson and Wolpert 2004; Tourville *et al.* 2008) that have been repurposed for the covert imitation and prediction of observed actions (i.e. mirror neurons (Gallese *et al.* 1996; Umiltà *et al.* 2001). But again, the evidence for such mechanisms is drawn from paradigms that are entirely individualistic (Watkins *et al.* 2003; Pulvermüller *et al.* 2006; Ito *et al.* 2009; Möttönen and Watkins 2009; Adank and Devlin 2010). Far from being irrelevant to our understanding of language, it seems that our understanding of interaction effects in language actually depends upon evidence gathered in non-interactive paradigms.

While the mechanisms underlying various alignment phenomena are present in non-interactive contexts, the case could be made that these mechanisms behave differently in social interactions. Branigan and colleagues (2007), for example, developed an interactive paradigm in which they were able to compare the rates of syntactic priming in participants in a conversational interaction with those in individuals who were merely side-participants. While they found syntactic priming effects in both groups, these effects were significantly stronger when a speaker had just been addressed than when he or she was merely listening to other individuals speak; but, as Branigan and colleagues themselves note, this effect is likely due to the fact that current addressee's were attending to the speaker more carefully than side-participants. Increased attention, of course, is not a uniquely interactive phenomenon.

This suggests that while alignment does increase in the context of conversational interactions, alignment is nevertheless explained by a host of mechanisms that do not operate only in interactive contexts.

There are some aspects of conversational alignment that are, in fact, uniquely interactive. For instance, Garrod and Pickering (2009) describe how participants in a dialogue also coordinate upon the timing of their utterances, which tends to yield fairly precise patterns of turn-taking (Ten Bosch *et al.* 2004; Levinson 2016). This phenomenon truly has no non-interactive equivalent, since turn-taking is by definition impossible in a monologue. We happily concede that this might be a case where an interactive context is necessary to truly grasp the nature of the phenomenon. However, Garrod and Pickering's explanation for our capacity for precise turn-taking in conversation invokes precisely the same covert imitation and priming mechanisms that explain other aspects of alignment. Thus, even if our knowledge of this phenomenon depends upon interactive experimental designs, we owe our understanding of it to individualistic research.

Thus, while dialogue is often cited as a paradigm case of an irreducibly interactive process, we would argue that conversational alignment arises from mechanisms that are not inherently interactive. In some cases, we do see these mechanisms operating differently in the context of interaction. In the case of turn-taking, we seem to have an instance of a genuine interaction effect. But other properties of dialogue, such as syntactic alignment, are also present in monologue; indeed, our very understanding of this aspect of dialogue is due to its study in non-interactive contexts.

4. How much does 'real' interaction matter?

It is sometimes suggested that interaction dynamics cannot be explained if we only look at the sum of the interactors' individual contributions to the encounter.[46] We don't wish to take a final stand on these issues in this chapter. In this section, we'd simply like to point out that most of the interactionists' own experiments seem to tacitly presuppose an individualist framework.

In a series of experiments, Schilbach and colleagues have investigated interaction-specific neural activation patterns of action-control (Schilbach *et al.* 2011), joint attention (Schilbach *et al.* 2010), and mutual gaze (Schilbach *et al.* 2006). In most of these experiments a subject is placed in an fMRI scanner engaging in some kind of interaction with a virtual character. Roughly, these experiments indicate that cues associated with interaction such as self-directed gaze are associated with differential neural activation in the medial prefrontal cortex, which is a region thought to be crucially implicated in social cognition (Van Overwalle 2009). For instance, Schilbach finds differentially increased neural activation in the medial prefrontal cortex for (a.) direct (vs. other-directed) gaze (Schilbach *et al.* 2006), and for following (vs. leading) someone's gaze (Schilbach *et al.* 2010). To account for the interactive element, all participants are made to believe that the virtual character is controlled by a real person with whom the interaction will subsequently take place. This belief, however, was false: the virtual character was entirely preprogrammed to

[46] For instance, De Jaegher *et al.* argue that 'interactive processes [...] complement and even replace individual mechanisms' (De Jaegher *et al.* 2010, 441). At the heart of this proposal is the idea that partitioning social cognitive processes into the cognitive mechanisms implemented by individual brains is unwarranted. Rather, it is the interaction between brains that should be considered explanatorily basic.

establish conditions of a controlled experiment. As a result, participants are not actually interacting. In terms of experimental design, this is fine; but what these experiments tacitly presuppose is that a subject's individual representation of a situation as interactive is sufficient to gain crucial insights in the cognitive significance of interaction.

One notable exception departing from the virtual-character paradigm is a study conducted by Cavallo and colleagues (2015). In this study, subjects established eye contact with a collaborator who was situated behind the fMRI scanner. The collaborator was visible to the participant via a mirror placed inside the scanner. In the experiment, either both subjects looked at each other (i.e. mutual gaze) or one of them looked away (i.e. averted gaze). In the control conditions participants either looked at their own eyes in a mirror reflection, or they looked at an image of the collaborator. Cavallo and colleagues found that mutual gaze differentially activates the anterior portions of the medial prefrontal cortex (mPFC).

As indicated above, Schilbach and colleagues found similar patterns of activation, even though they relied upon paradigms that used virtual characters (Schilbach *et al.* 2006b, 2010b). Comparing these experiments, it seems that real interaction does not seem to have made a crucial difference to activation in the mPFC, which was the main finding in the mutual gaze condition. Furthermore, Cavallo and colleagues found that neural activation was independent of whether subjects actually established eye contact or whether subjects merely knew that the collaborator was looking at them. Hence, it was the 'mere belief of being seen' (Cavallo *et al*. 2015, 67) which

accounted for the distinct pattern of neural activation; actual interaction seemed irrelevant. Importantly, while experiments by Schilbach *et al*. support the idea that even simulated interaction leads to activation in the mPFC, the study by Cavallo *et al*. provides direct comparative evidence for the claim that real interaction is not crucial for the relevant neural activation patterns to occur. Lastly, while Schilbach also reports increased activity in the amygdala, Cavallo finds no such activity.[47] And even if differential activation in the amygdala were to indicate a difference between virtual and real interactions, the absence of such activity in a real interactive conditions is rather bad news for the interactionists, who have pointed out that emotional engagement is a crucial cognitive element in social interactions (Reddy 2008; Schilbach *et al.* 2013).

Together these observations suggest that, at least in gaze paradigms, it is more significant whether a subject believes that she engaged in an interaction; and not so much whether she is actually engaged in an interaction.

5. *Conclusion*

Our aim in this chapter has been to draw attention to the various conceptual and methodological confusions that arise when we over-emphasize the notion of interaction in social cognition research. First, we argued that De Jaegher and colleagues' prominent definition of interaction diverged significantly from the intuitive consensus, and also seems to equivocate on the notion of autonomy. Second,

[47] Notably, involvement of the amygdala has been inconsistent throughout an array of studies investigating mutual gaze. For instance, while a number of authors (Kawashima *et al.* 1999; Wicker *et al.* 2003; Sato *et al*. 2004; Schilbach *et al.* 2006) have found activation in the amygdala during mutual gaze, several others have not (Calder *et al.* 2002; Pageler *et al.* 2003).

we illustrated how interactive paradigms potentially confound genuine interaction effects with the effects of factors that merely co-occur with interaction. Finally, we showed that genuine interactions are not needed to study the effects of interaction on cognition: the mere representation of interactivity will often do just as well. Genuine interactivity, although often the distal cause of such representations, do not play a special role in explaining these effects.

However, our goal in this chapter is not completely negative, and we are not wholly opposed to interactive experimental designs; rather, we advocate for a complementary, multi-method approach that includes both interactive and non-interactive methods. However, when interactive designs are used, we advise that researchers remain cautious in their interpretations, and that they implement appropriate controls before attributing the effects they discover to interaction as such. We hope that by drawing attention to the various confounds and confusions that arise in interactive experimental designs, we have clarified the significance of interaction in social cognition research. With this added clarity, we hope, researchers will now be better positioned to pursue the goal of making experimental paradigms in social cognition research more ecologically valid. With this end in mind, we have three general suggestions for future research:

1. Interaction is complicated, but defining it doesn't have to be: While the philosophical debate surrounding the ontology of social interaction is still ongoing, this debate need not impinge upon practical applications of the concept of interaction in research contexts. The notion of autonomy, in

particular, serves merely to obscure, rather than to clarify, the meaning of 'social interaction'. In lieu of the one provided by De Jaegher and colleagues, we have offered our own definition that captures the intuitive notion of social interaction with minimal conceptual baggage.

2. Interaction effects versus social effects on social cognition: Ordinary social interactions are complex events, which tend to involve a cluster of social elements that are not themselves interactive. This makes it difficult to study the effects of interaction as such, because we must distinguish the effects of interaction from concomitant social factors. Researchers interested in improving upon the ecological validity of social cognition paradigms must recognize these factors could potentially dissociate from interaction, and ought to be investigated in their own right.

3. Real versus represented interaction: Many of the purported effects of interaction on social cognition can also be found in pseudo-interactive paradigms. This shows that paradigms manipulating beliefs about interaction can be just as informative as the paradigms that involve genuine interaction. Once this individualist insight into the 'interactionist turn' is taken on board, it opens up practical possibilities for social cognition research by making the problem of social interaction more empirically tractable.

Bibliography

Adank, P., Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, 49, 1124–32.

Apperly, I. A., Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological review*, 116(4), 953.

Auvray, M., Rohde, M. (2012). Perceptual crossing: the simplest online paradigm. *Frontiers in human neuroscience*, *6*, 181.

Auvray, M., C. Lenay, J. Stewart. (2009). Perceptual interactions in a minimalist virtual environment. *New ideas in psychology,* 27(1), 32–47.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 3–8.

Baillargeon, R. L., Baillargeon, R., Southgate, V. (2018). Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*.

Baillargeon, R., R. Scott, Z. H. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences,* 14(3), 110–118.

Barwise, J. (1988). Three views of common knowledge. Pages 365–379 of: *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann Publishers Inc.

Bergmann, M. (2005). Defeaters and higher-level requirements. *The Philosophical Quarterly, 55*, 419–436.

Bicchieri, C. (2006). *The Grammar of Society*, Cambridge: Cambridge University Press.

Binmore, K., Samuelson, L. (2001). Coordinated action in the electronic mail game. *Games and Economic Behavior*, 35(1-2), 6–30.

Blomberg, O. (2015). Shared Goals and Development. *Philosophical Quarterly,* 65 (258), 94-101.

Blomberg, O. (2016). Common knowledge and reductionism about shared agency. *Australasian Journal of Philosophy*, 94 (2), 315–326.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18, 355–87.

Bock, K., Dell, G. S., Chang, F., Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production, *Cognition*, 104, 437–58.

Boesch, C. (1994). Cooperative hunting in wild chimpanzees. *Animal Behaviour* 48(3), 653–667.

Boesch, C., Boesch-Acherman, H. (2000). *The chimpanzees of the Tai forest: Behavioural ecology and evolution*. Oxford: Oxford University Press.

Bonanno, G. (1996). On the logic of common belief. *Mathematical Logic Quarterly*, *42*(1), 305–311.

Branigan, H. P., Pickering, M.J., Cleland, A.A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75, 13–25.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information.

Bratman, M. E. (1992). Shared cooperative activity. *Philosophical Review,* 101 (2), 327-341.

Bratman, M. E. (1993). Shared intention. *Ethics*, 104(1), 97-113.

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. New York: Oxford University Press.

Bratman, M. E. (2003). Autonomy and hierarchy. *Social Philosophy and Policy*, 20, 156–76.

Bratman, M. E. (2007). *Structures of agency: Essays*, New York, NY: Oxford University Press.

Brugger, A., Lariviere, L. A., Mumme, D. L., Bushnell, E. W. (2007). Doing the Right Thing: Infants' Selection of Actions to Imitate From Observed Event Sequences. *Child Development*, 78, 806–24.

Bullinger A. F., Wyman E, Melis, A. P., Tomasello M. (2011). Coordination of chimpanzees (Pan troglodytes) in a stag hunt game. *International Journal of Primatology,* 32, 1296-1310.

Burge, T. (1975). On Knowledge and Convention. *Philosophical Review*, 84, 249–255.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One*, 12(4).

Buttelmann, D., Carpenter, M., Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337-342.

Buttelmann, D., Carpenter, M., Call, J., Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science*, 10, F31–F38 .

Butterfill, S. (2011). Joint action and development. *The Philosophical Quarterly*, *62*(246), 23-47.

Butterfill, S. A., Apperly, I. A. (2013a). How to construct a minimal theory of mind. *Mind & Language*, 28(5), 606-637.

Butterfill, S. A. (2013b). Interacting mindreaders. *Philosophical Studies,* 165(3), 841–863.

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., and Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, 40, 1129–38.

Call, J., Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences,* 12, 187–192.

Camerer, C. F. Ho, T.-H. Chong, J. K. (2003). Models of Thinking, Learning and Teaching in Games. *American Economic Review*, 93, May, 192–5.

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and brain sciences,* 32(02), 121–138.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Carruthers, P. (2013). Mindreading in Infancy. *Mind and Language*, 28 (2),141-172.

Carruthers, P. (2015). Perceiving mental states. *Consciousness and cognition,* 36, 498–507.

Cavallo, A., Lungu, O., Becchio, C., Ansuini, C., Rustichini, A., Fadiga, L. (2015). When gaze opens the channel for communication: Integrative role of IFG and MPFC. *NeuroImage*, 119, 63–9.

Clark, H. H., Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A. K., Webber, B., Sag, I. (eds), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

Craft, J. L., Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology*, 83, 415–20.

Csibra, G., Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance XXI*, 249–74.

Csibra, G., Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, 13, 148–53.

Csibra, G., Gergely, G. (2006). Social learning and social cognition: the case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance,* XXI, 249–274.

Cubitt, R. P., Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of David Lewis' game theory. *Economics and Philosophy*, 19 (02), 175–210.

Dale, R., Kirkham, N.Z., Richardson, D.C. (2011). The dynamics of reference and shared visual attention. *Frontiers in Psychology*, 2, 1–11.

Daoutis, C. A., Franklin, Riddett, A., Davies, C. (2006). Categorical effects in children's colour search: a cross-linguistic comparison. *British Journal of Developmental Psychology,* 24(2), 373–400.

Davidson, D. (1978). Intending. In *Philosophy of history and action* (pp. 41–60). Springer.

Davidson, P. R., Wolpert, D. M. (2004). Internal models underlying grasp can be additively combined. *Experimental Brain Research*, 155, 334–40.

Davies, M. (1987). Relevance and Mutual Knowledge. *Behavioral and Brain Sciences,* 10/4: 716-17.

De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the cognitive sciences,* 6(4), 485–507.

De Jaegher, H., Di Paolo, E., Gallagher, S. (2010). Can social interaction constitute social cognition?. *Trends in cognitive sciences,* 14(10), 441–447.

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain sciences*, 1(4), 568-570.

Dienes, Z., Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22, 735–808.

Elekes, F., Varga, M., Király, I. (2016). Evidence for spontaneous level-2 perspective taking in adults. *Consciousness and Cognition*, 41, 93–103.

Fagin, Ronald, Halpern, Joseph Y., Moses, Yoram, & Vardi, Moshe Y. 1995. Reasoning about Knowledge. MIT Press.

Flavell, J. H., Everett, B. a., Croft, K., Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology*, 17, 99–103.

Freundlieb, M., Kovács, Á. M., Sebanz, N. (2016). When do humans spontaneously adopt another's visuospatial perspective?. *Journal of experimental psychology: human perception and performance*, *42*(3), 401.

Friedman, J. (2013). Suspended judgment. *Philosophical Studies,* 162 (2), 165–181.

Friedman, J. (2013b). Rational Agnosticism and Degrees of Belief. *Oxford Studies in Epistemology,* 4:57.

Friedman, J. (2017). Why suspend judging?. *Nous*, 51 (2), 302–326.

Fusaroli, R., Bjørndahl, J. S., Roepstorff, A., Tylén, K. (2016). A heart for interaction: Shared physiological dynamics and behavioral coordination in a collective, creative construction task. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 1297–310.

Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction?. *Journal of Consciousness Studies*, 8, 83–108.

Gallagher, S. (2009). Two problems of intersubjectivity. *Journal of Consciousness Studies*, 16, 289–308.

Gallagher, S. (2008a). Inference or interaction: social cognition without precursors. *Philosophical Explorations,* 11(3), 163–174.

Gallagher, S., Hutto, D. (2008b). Understanding others through primary interaction and narrative practice, *The shared mind: Perspectives on intersubjectivity*, Philadelphia, PA: John Benjamins Publishing Compary, 17–38.

Gallagher, S., Povinelli, D. J. (2012). Enactive and Behavioral Abstraction Accounts of Social Understanding in Chimpanzees, Infants, and Adults. *Review of Philosophy and Psychology*, 3, 145–69.

Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593–609.

Gallotti, M., Frith, C.D. (2013). Social cognition in the we-mode. *Trends in Cognitive Sciences*, 17, 160–5.

Garrod, S., Pickering, M.J. (2004). Why is conversation so easy?. *Trends in Cognitive Sciences*, 8, 8–11.

Garrod, S., Pickering, M.J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, 1, 292–304.

Gergely, G. (2010). Kinds of agents. *The Wiley-Blackwell book of Childhood Cognitive Development,* 22, 76.

Gergely, G., Csibra, G. (2005). The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy. *Interaction Studies*, 6, 463–81.

Gilbert, M. (1981). *On Social Facts*, New York: Routledge.

Gilbert, M. (2003). The structure of the social atom: Joint commitment as the foundation of human social behavior. In F. Schmitt (Ed.). *Socializing metaphysics* (pp. 39–64). Lanham, MD: Rowan & Littlefield.

Gray, Jim. 1978. Notes on Data Base Operating Systems. Operating Systems, An Advanced Course. (Pages 393–481). Springer-Verlag

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66, 377–88.

Grice, H. P. (1989). *Studies in the Way of Words*, Cambridge, Mass.: Harvard University Press.

Guagnano, D., Rusconi, E., Umiltà, C.A. (2010). Sharing a task or sharing space? On the effect of the confederate in action coding in a detection task. *Cognition*, 114, 348–55.

Hájek, A. (1998). Agnosticism Meets Bayesianism. *Analysis*, *58*(3), 199-206.

Hájek, A. (2016). Deliberation welcomes prediction. *Episteme*, 13 (4), 507–528.

Hamann K., Warneken F., Greenberg J. R., Tomasello M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature,* 476,328-331

Hare, B., Call, J., Tomasello, M. (2001). Do chimpanzees know what conspecifics know?. *Animal Behavior,* 61, 139–151.

Hare, B., Call, J., Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101(3), 495-514.

Hare, B., Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, 68(3), 571-581.

Harman, G. (1977). Review of Jonathan Bennett's Linguistic Behaviour. *Language*, 53(2), 417–424.

Heal, J. (1978). Common knowledge. *The Philosophical Quarterly*, 28(111), 116–131.

Hédoin, C. (2014). A framework for community-based salience: common knowledge, common understanding and community membership. *Economics & Philosophy*, *30*(3), 365-395.

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton; Oxford: Princeton University Press.

Herrmann, E., Call, M. V., Hernandez-Lloreda, B., Hare, B, Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science,* 317 (5843),1360–1366.

Herrmann, E., Hernandez-Lloreda, J., Call, B., Hare, Tomasello, M. (2010). The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science,* 21 (1),102–110.

Herschbach, M. (2012). On the role of social interaction in social cognition: a mechanistic alternative to enactivism. *Phenomenology and the Cognitive Sciences,* 11(4), 467–486.

Heyes, C. (2014). False belief in infancy: a fresh look. *Developmental science*, 17(5), 647-659.

Horty, J. F. (2012). *Reasons as defaults*. Oxford University Press.

Hutto, D. D. (2009). Folk psychology as narrative practice. *Journal of Consciousness Studies,* 16(6–8): 9–39.

Hutto, D. D. (2004). The limits of spectatorial folk psychology. *Mind & Language*, *19*(5), 548-573.

Hutto, D. D. (2007). The narrative practice hypothesis: origins and applications of folk psychology. *Royal Institute of Philosophy Supplement*, 60, 43–68.

Ito, T., Tiede, M., Ostry, D.J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 1245–8.

Kaminski J., Call J., Tomasello M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109,224-234.

Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Ito, K., Fukuda, H., Kojima, S., and Nakamura, K. (1999). The human amygdala plays an important role in gaze monitoring. *Brain*, 122, 779–83.

Kneeland, T. (2012). Coordination under limited depth of reasoning. University of British Columbia Working Paper.

Köster, M., Ohmer, X., Nguyen, T. D., Kärtner, J. (2016). Infants understand others' needs. *Psychological science*, 27(4), 542-548.

Kovács, Á. M., Téglás, E., Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830-1834.

Krachun C., Carpenter M., Call J., Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, 12,521-535.

Krupenye, C., Kano, F., Hirata, S., Call, J., Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308), 110-114.

Kutz, C. (2000). Acting Together. *Philosophy and Phenomenological Research,* 61 (1), 1–31.

Laurence, B. (2011). An Anscombian approach to collective action. In Anton Ford, Jennifer Hornsby & Frederick Stoutland (eds.), *Essays on Anscombe's Intention*. Harvard University Press.

Lederman, H. (2017). Uncommon knowledge. *Mind*, fzw072. doi: 10.1093/ mind/fzw072

Lederman, H. (2018). Two paradoxes of common knowledge: Coordinated attack and electronic mail. *Noûs*, 52(4), 921-945.

Levinson, S. C. (2016). Turn-taking in Human Communication - Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, 20, 6–14.

Lewis, D. K. (1969). *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.

Lismont, L., Mongin, P. (2003). Strong completeness theorems for weak logics of common belief. *Journal of Philosophical Logic*, 32(2), 115–137.

Liszkowski, U., Carpenter, M., Striano, T., Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, 7(2), 173-187.

Masangkay, Z. S., McCluskey, K. a, McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., and Flavell, J. H. (1974). 'The early development of inferences about the visual percepts of others.', *Child development*, 45, 357–66.

Matheson, H., Moore, C., Akhtar, N. (2013). The development of social learning in interactive and observational contexts, *Journal of Experimental Child Psychology*, 114, 161–72.

Melis, A. P., Hare, B., Tomasello, M. (2006). Engineering cooperation in chimpanzees: tolerance constraints on cooperation. *Animal Behavior,* 72, 275-286.

Mertens, J., Zamir, S. (1985). Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 14(1), 1–29.

Michael, J. (2011). Interactionism and mindreading. *Review of Philosophy and Psychology,* 2(3), 559–578.

Michael, J., Pacherie, E. (2014). On commitments and other uncertainty reduction tools. *Journal of Social Ontology*, 1–34.

Michelon, P., Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & psychophysics*, 68, 327–37.

Milgrom, P. (1981). An axiomatic characterization of common knowledge. Econometrica. *Journal of the Econometric Society*, 219–222.

Miller, S. (2001). *Social action: A teleological account*. Cambridge University Press.

Moll, H., Carpenter, M, Tomasello, M. (2011a). Social engagement leads 2-year-olds to overestimate others knowledge. *Infancy,* 16(3), 248–265.

Moll, H., Meltzoff, A. (2011). Perspective-taking and its foundation in joint attention", in: J. Roessler (Ed.), *Perception, Causation, and Objectivity. Issues in Philosophy and Psychology* (Oxford: Oxford University Press).

Moore, R. E. (1979). Refraining. *Philosophical Studies, 36*, 407–424.

Moore, R. (2013). Imitation and conventional communication. *Biology and Philosophy*, 28 (3), 481–500.

Möttönen, R., Watkins, K. E. (2009). Motor Representations of Articulators Contribute to Categorical Perception of Speech Sounds. *Journal of Neuroscience*, 29, 9819–25.

Murray, L., Trevarthen, C. (1985). The infant in mother-infant communication. *Journal of Child Language*, 13, 15–29.

Nagel, J. (2013). Knowledge as a mental state. *Oxford Studies in Epistemology*, 4, 273–306.

Noe, A. (2008). Précis of "Action in Perception: Philosophy and Phenomenological Research" *Philosophy and Phenomenological Research,76*(3), 660-665.

O'Hear, A. (ed) (2009). *Epistemology*. Cambridge University Press.

Onishi, K. H., Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *Science*, 308(5719), 255-258.

Pageler, N. M., Menon, V., Merin, N. M., Eliez, S., Brown, W. E., Reiss, A. L. (2003). 'Effect of head orientation on gaze processing in fusiform gyrus and superior temporal sulcus', *NeuroImage*, 20, 318–29.

Paternotte, C. (2011). Being realistic about common knowledge: a Lewisian approach. *Synthese*, *183*(2), 249-276.

Penn, D. C., Povinelli, D. J. (2013). 3 The Comparative Delusion: The "Behavioristic/Mentalistic" Dichotomy in Comparative Theory. *Agency and joint attention*, 62.

Pickering, M. J., Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, 3, 136–41.

Pickering, M. J., Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27, 169-190-226.

Pickering, M. J., Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and brain sciences*, 36, 329–47.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences,* 1(4), 515-526.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 7865–70.

Qureshi, A.W., Apperly, I., Samson, D. (2010). 'Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: evidence from a dual-task study of adults.', *Cognition*, 117, 230–6.

Redcay, E., Rice, D., Saxe, R. (2013). Interaction versus observation: a finer look at this distinction and its importance to autism. *Behavioral and Brain Sciences,* 36(04), 435–435.

Reddy, V. (2008). *How infants know minds*. Cambridge: Harvard University Press.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *The American Economic Review*, 385–391.

Samson, D., Apperly, I., Braithwaite, J. J., Andrews, B. J., Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see, *Journal of Experimental Psychology: Human Perception and Performance*, 36, 1255–66.

Sato, W., Yoshikawa, S., Kochiyama, T., Matsumura, M. (2004). The amygdala processes the emotional significance of facial expressions: an fMRI investigation using the interaction between expression and face direction. *NeuroImage*, 22, 1006–18.

Schilbach, L. (2014). On the relationship of online and offline social cognition. *Frontiers in human neuroscience,* 8.

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences,* 3, 130–135.

Schilbach, L., Eickhoff, S. B., Cieslik, E. C., Kuzmanovic, B., Vogeley, K. (2012). Shall we do this together? Social gaze influences action control in a comparison group, but not in individuals with high-functioning autism. *Autism*, *16*(2), 151-162.

Schilbach, L., Eickhoff, S. B., Cieslik, E., Shah, N. J., Fink, G. R., Vogeley, K. (2011). Eyes on me: an fMRI study of the effects of social gaze on action control, *Social Cognitive and Affective Neuroscience*, 6, 393–403.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., (2013A). Toward a second person neuroscience. *Behavioral and Brain Sciences,* 36(04), 393–414.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., Vogeley, K. (2013B). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36, 393–414.

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *Journal of Cognitive Neuroscience,* 22(12), 2702–2715.

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia,* 44(5), 718–730.

Schmid, H. B. (2016). On knowing what We're doing together: Groundless group self-knowledge and plural self-blindness. *The epistemic life of groups: essays in the epistemology of collectives. OUP, Oxford*, 51-74.

Schneider, D., Nott, Z. E., Dux, P. E., (2014). Task instructions and implicit theory of mind. *Cognition,* 133(1), 43–47.

Schönherr, J. (2017). What's so Special About Interaction in Social Cognition?. Review of Philosophy and Psychology, 8(2), 181-198.

Schönherr, J. (2018). Lucky joint action, *Philosophical Psychology*, 32:1, 123-142.

Schönherr, J., Westra, E. (2019). Beyond 'interaction': How to understand social effects on social cognition. *British Journal for the Philosophy of Science*, 70(1), 27-52.

Searle, J. R. (1990). Collective Intentions and actions. *Intention in communication*, *401*, 401.

Sebanz, N., Knoblich, G., Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition*, 88, 11–21.

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, 22(2), 353-360

Senju, A., Southgate, V., Snape, C., Leonard, M., Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological science*, 22(7), 878-880.

Shimpi, P. M., Akhtar, N., Moore, C. (2013). Toddlers imitative learning in interactive and observational contexts: the role of age and familiarity of the model. *Journal of experimental child psychology,* 116(2), 309–323.

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro J., Schapiro, S. J. (2005). Chimpanzees are indifferent to the welfare of unrelated group members. *Nature*, 437,1357-1359

Sillari, G. (2008). Common knowledge and convention. *Topoi*, 27 (1-2), 29–39.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Southgate, V., Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition*, 130(1), 1-10.

Southgate, V., Senju, A., Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18(7), 587-592.

Spaulding, S. (2010). Embodied cognition and mindreading. *Mind & Language,* 25(1), 119–140.

Sturgeon, S. (2010). Confidence and coarse-grained attitudes. In *Oxford studies in Epistemology* (Vol. 3). Oxford: Oxford University Press.

Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology*, 1 (1), 143–166.

Surian, L., Caldi, S., Sperber, D. (2007). Attribution of beliefs by 13-month-oldinfants. *Psychological Science,* 18(7), 580–586.

Surtees, A., Apperly, I. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development,* 83, 452–460.

Surtees, A., Butterfill, S., and Apperly, I. (2012). Direct and indirect measures of Level-2 perspective-taking in children and adults. *The British journal of developmental psychology*, 30, 75–86.

Surtees, A., Samson, D., Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, 148, 97–105.

Tal, E., Comesana, J. (2016). Is evidence of evidence evidence?. *Nous*, 50 (4).

Ten Bosch, L., Oostdijk, N., De Ruiter, J. P. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. *International Conference on Text, Speech and Dialogue*, Berlin: Springer, 563–70.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on pre-attentive color perception. *Proceedings of the National Academy of Sciences,* 106(11), 4567–4570.

Tomasello, M. (2011). Human culture in evolutionary perspective. In M. Gelfand (Ed.), *Advances in Culture and Psychology*. Oxford U. Press. [pdf]

Tomasello, M. (2016). *A natural history of human morality.* Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Tourville, J. A., Reilly, K. J., Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, 39, 1429–43.

Tuomela, R. (1974). *Human Action and its Explanation*. Institute of Philosophy, University of Helsinki.

Tuomela, R. (2005). We-intentions revisited. *Philosophical Studies*, 125(3), 327–369.

Tylén, K., Allen, M., Hunter, B. K., Roepstorff, A. (2012). Interaction vs. observation: distinctive modes of social cognition in human brain and behavior? A combined fMRI and eye-tracking study. *Frontiers in human neuroscience*, 6.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., Rizzolatti, G. (2001). I Know What You Are Doing. *Neuron*, 31, 155–65.

Van de Vondervoort, J. W., Hamlin, J. K. (2017). Preschoolers' social and moral judgments of third-party helpers and hinderers align with infants' social evaluations. *Journal of experimental child psychology*, 164, 136-151.

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30, 829–58.

Vanderschraaf, P., Sillari, G. (2014). Common knowledge. In E. N. Zalta (Ed.), The stanford encyclopedia of philosophy (Spring 2014 ed.). Metaphysics Research Lab, Stanford University.

Vesper, C., Butterfill, S., Knoblich, G., Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*(8-9), 998-1003.

Vlainic, E., Liepelt, R., Colzato, L.S., Prinz, W., Hommel, B. (2010). The virtual co-actor: The social Simon effect does not rely on online feedback from the other. *Frontiers in Psychology*, 1, 1–6.

Wang, L., Leslie, A. M. (2016). Is implicit theory of mind the 'Real Deal'? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language*, 31(2), 147-176.

Warneken, F., Chen, F., Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child development*, *77*(3), 640-663.

Warneken, F., Lohse, K., Melis, A. P., Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological science*, 22(2), 267-273.

Watkins, K. E., Strafella, A.P., Paus, T. (2003). 'Seeing and hearing speech excites the motor system involved in speech production', *Neuropsychologia*, 41, 989–94.

Wedgwood, R. (2012). The Aim of Belief. *Philosophical Perspectives*, 36 (s16), 267 – 297.

Wegner, D. (2002). The Illusion of Conscious Will. MIT Press.

Wellman, H. M., Cross, D., Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3), 655-684.

Wellman, H. M., Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75, 523–541.

Wicker, B., Perrett, D.I., Baron-Cohen, S., Decety, J. (2003). Being the target of another's emotion: a PET study. *Neuropsychologia*, 41, 139–46.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, E, Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences,* 104(19), 7780–7785.

Woodward, A. L., Sommerville, J. A., Gerson, A., Henderson, A. M., Buresh, J. (2009). The emergence of intention attribution in infancy. *Psychology of learning and motivation,* 51, 187–222.

Yoon, J. M., Johnson, M. H., Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences,* 105(36), 13690–13695.