Matti Sarkia

A Model-Based and Mechanistic Approach to Social Coordination

Academic dissertation

To be presented, with the permission of the Faculty of Social Sciences
of the University of Helsinki, for public examination in
Room 302, Athena-building, on Tuesday, 21 June 2022,
at 10 am.

Filosofisia tutkimuksia Helsingin yliopistosta
Filosofiska Studier från Helsingfors universitet
Philosophical Studies from the University of Helsinki

Publishers:
Theoretical Philosophy
Social and Moral Philosophy
Philosophy (Swedish)

P.O. Box 24 (Unioninkatu 40A)
00014 University of Helsinki
Finland

Editors:
Michiru Nagatsu
Samuli Reijula
Thomas Wallgren

ISBN 978-951-51-8238-8 (paperback)
ISBN 978-951-51-8239-5 (PDF)
ISSN 1458-8331 (series)
Unigrafia, Helsinki 2022.

Abstract

This dissertation deals with topics of interdisciplinary exchange and integration that arise at the intersection of analytic philosophy, behavioral and cognitive science, and the social sciences. Thematically, most of this dissertation is concerned with the phenomenon of social coordination, how it is studied in different scientific fields, and how their theoretical insights can be brought together and negotiated with one another. Methodologically, I draw on approaches in contemporary philosophy of science that are related to theoretical modeling and mechanistic explanation, understood as distinctive methodological strategies that scientists use to study complex phenomena. My focus is on studies of *shared intentionality* (which serves as a prerequisite for many central forms of social coordination) in analytic philosophy, and how philosophical studies of shared intentionality relate to studies of social coordination in other scientific disciplines, such as cognitive science, developmental psychology, evolutionary anthropology, and economics. In this respect, I compare the methodological status of philosophical studies of shared intentionality to the methodological status of theoretical models in science, and the activity of conceptual analysis to the activities of scientific modeling and model-construction (articles I-III of my dissertation). Moreover, I argue that the mechanistic approach to explanation can play an important role in bringing together different disciplinary perspectives on social coordination (articles IV-VI of my dissertation).

Preface

This dissertation could not have been completed without the help and support of many colleagues, friends, and family. The philosophy discipline in Helsinki has been an exceptionally supportive community to work in over the years. My immediate and extended families have been a supportive force throughout my research career. They and numerous friends, as well as my one-year old son, have provided much needed digressions from work and reminded me that there are more important things in life than the most recent academic debates. Here I have singled out some individuals, who had an especially important influence on my life while writing this dissertation.

Professor Raimo Tuomela was my first academic supervisor, who was an encouraging presence and constant source of inspiration in the early stages of my dissertation research and before, when I was working as his research assistant during my master's studies. Pekka Mäkelä played an important role in introducing me to both Raimo and Professor Seumas Miller, with whom I was able to work for a brief period of time at the University of Melbourne on a topic relating to the philosophy of social institutions. Pekka and other members of Raimo's social action research group, especially Raul Hakli and Kaarlo Miller, were the source of many incisive ideas and innumerable laughs at our regular research seminars. Arto Laitinen, Mikko Salmela, and Maj Tuomela deserve equal thanks for participating in these meetings and bringing them up to the highest standards of philosophical argumentation.

As I was beginning my doctoral research, I found the social action research group integrated into the Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences, which was directed by Professor Uskali Mäki and Professor Petri Ylikoski (as vice-director). Uskali was a fair-minded director with a deep concern for issues of procedure, and his research on economic modeling gave me the impetus to take my doctoral research in a new direction. Petri crosses disciplinary

boundaries with ease and is impressive in the breadth of his knowledge, knowing everything about almost anything else that you might ask him about. Many of the brilliant post docs and advanced researchers that were recruited to the CoE were outstanding academic role models for an aspiring researcher, and several of them have by now advanced to (well-deserved) more senior academic positions. Some individuals from this group whom I would like to thank separately include (in alphabetic, hence no particular order) Emrah Aydinonat, Alessandra Basso, Alkistis Elliot-Graves (as well as her brilliant and genuine partner Yiannis Kokosalakis), Marion Godman, Harold Kincaid, Saana Jukola, Tuukka Kaidesoja, Tomi Kokkonen, Inkeri Koskinen, Jaakko Kuorikoski, Caterina Marchionni, Carl Martini, Luis Mireles-Flores, Michiru Nagatsu, Päivi Seppälä, Samuli Reijula, Anita Välikangas, and Julie Zahle (apologies for those whose names I have forgotten to mention).

During the past few years of my doctoral research, I have had the privilege of working with Doctor Tuukka Kaidesoja in two research projects on the cognitive social sciences that have been funded by the Emil Aaltonen Foundation (2019-2021) and the Kone Foundation (2021-2024). Tuukka convinced me that the mechanistic approach to explanation can play a major role in theorizing relations between the cognitive sciences and the social sciences, and he encouraged me to broaden my research focus from philosophical debates relating to shared intentionality to more general issues concerning the interdisciplinary study of social coordination (as well as other topics). Tuukka has been an exceptional team leader, who cares deeply about the junior members of his research team, distributes resources unselfishly, and always does more than his own fair share in joint projects. I consider myself extremely lucky to be part of his research group.

Apart from my close colleagues in the philosophy discipline, I would like to extend my thanks to Mrs. Karoliina Kokko and Mrs. Terhi Mölsä at the Fulbright Finland Foundation, which enabled me to carry out a year of research at the City University of New York (CUNY) Graduate

Center under the supervision of Distinguished Professor Peter Godfrey-Smith. Apart from serious one-on-one feedback sessions with Peter and the excellent seminars offered by CUNY, I was able to reconnect with my academic friend Bradley Turner, with whom I had the good fortune of sharing an office in Helsinki while he was out on his own Fulbright scholarship a few years earlier. In addition to Bradley, many other friends have been an important part of my personal and academic life while writing this dissertation—you know who you are, so I won't name any names.

Last, I would like to thank my immediate and extended families for their support throughout my academic career: especially my father Martti for encouraging me to pursue my own path in life, my brother Erkki for being such a beacon of academic excellence, Sampsa and Olli for their witticisms and unrelenting intellectual curiosity, and my partner Katri for believing that philosophical research has more to contribute to society than tedious arguments that are impossible to explain to the uninitiated. However, the greatest thanks go to my 14-month old son, Sami, who has learned to walk and to utter his first words during the months when I have been writing this introductory essay. My achievements in this dissertation pale in comparison to the intellectual barriers that he has already broken through during his so far brief, but vigorous existence. When he learns to talk, he will no doubt have much to teach us.

In Helsinki on May 25, 2022,

Matti Sarkia

List of original publications

This thesis is based on the following publications:

I **Modeling intentional agency: a neo-Gricean framework**
Matti Sarkia (2021)
Synthese, 199, 7003-7030.

II **A family of models of shared intentionality**
Matti Sarkia (2022)
Under review

III **A model-based approach to social ontology**
Matti Sarkia (2021)
Philosophy of the Social Sciences. OnlineFirst.

IV **Minimalism and maximalism in the study of shared intentional action**
Matti Sarkia (2016)
Philosophy of the Social Sciences, 46(2) 168-188.

V **Mechanistic explanation, interdisciplinary integration, and interpersonal social coordination**
Matti Sarkia (2022)
Under review

VI **Mechanistic explanations in the cognitive social sciences: lessons from three case studies**
Matti Sarkia, Tuukka Kaidesoja, and Mikko Hyyryläinen (2020)
Social Science Information, 59(4) 580-603

The publications are referred to in the text by their roman numerals.

Part I: Introductory essay

1. Introduction

This dissertation deals with topics of interdisciplinary exchange and integration that arise at the intersection of analytic philosophy, behavioral and cognitive science, and the social sciences. Thematically, most of this dissertation is concerned with the phenomenon of social coordination, how it is studied in different scientific fields, and how their theoretical insights can be brought together and negotiated with one another. Methodologically, I draw on approaches in contemporary philosophy of science that are related to theoretical modeling and mechanistic explanation, understood as distinctive methodological strategies that scientists use to study complex phenomena. My focus is on studies of *shared intentionality* (which serves as a prerequisite for many central forms of social coordination—more on this below) in analytic philosophy, and how philosophical studies of shared intentionality relate to studies of social coordination in other scientific disciplines, such as cognitive science, developmental psychology, evolutionary anthropology, and economics. In this respect, I compare the methodological status of philosophical studies of shared intentionality to the methodological status of theoretical models in science, and the activity of conceptual analysis to the activities of scientific modeling and model-construction (articles I-III of my dissertation). Moreover, I argue that the mechanistic approach to explanation can play an important role in bringing together different disciplinary perspectives on social coordination (articles IV-VI of my dissertation). This introductory essay leads into my subject matter, provides background to the general methodological principles that motivate my study, and motivates some of its central claims. Thus I will not seek to systematically reiterate the claims that I make in my dissertation—rather, the reader is referred to the original articles for detailed discussion and arguments.

This introductory essay is divided into five chapters. The second chapter begins with a general introduction to the subject matter of my dissertation, shared intentionality and social coordination (sections 2a-2b). In addition to perspectives from philosophy and game theory (sections 2c and 2e), it overviews research on social coordination in developmental psychology and evolutionary anthropology (section 2d). The third chapter turns to my methodological outlook, first discussing interdisciplinarity as a constraint on theory formation in science (3a), then the methodology of model-construction as indirect representation of the world by the mediation of a surrogate system (3b), and the frameworks of mechanistic explanation and mechanism discovery (3c). The fourth chapter motivates (but does not defend in detail—for that, see the original articles) the central ideas that I advance in my dissertation, beginning with a defense of methodological naturalism as a perspective on philosophical practice (4a), then presenting my views about conceptual analysis as a form of model-construction (4b), and the framework of mechanistic explanation as a template for division of labor between different scientific disciplines studying social coordination (4c). The fifth and concluding chapter provides some comments (which are not intended to serve as a summary or to replace an abstract) to the original research articles that make up the bulk of my dissertation.

2. Social coordination as a topic of interdisciplinary investigation

This part of my introductory essay leads into the subject matter of my dissertation, social coordination (in a broad, interdisciplinary sense) and shared intentionality (as it is commonly understood in philosophical parlance). In section 2a, I motivate my study of social coordination by the "ultra-sociality" of the human species, and discuss how the human capacity for social coordination has become a central topic of

investigation in many different scientific disciplines. Moreover, I distinguish planned coordination from emergent coordination, and discuss the relation between social coordination and social cognition, including the capacity for meta-representation. In section 2b, I introduce the terminology of individual and shared intentionality, distinguish shared intentional actions from unintentionally coordinated behaviors, and argue that some, but not all forms of social coordination depend on shared intentionality. In section 2c, I digress on some well-known philosophical accounts of shared intentionality, including work by Michael Bratman, Margaret Gilbert, John Searle, and Raimo Tuomela. In section 2d, I return to a broader, interdisciplinary outlook to discuss the ontogeny and phylogeny of human capacities social coordination, and what constraints they set on philosophical theorizing about shared intentionality. In section 2e, I discuss the ways in which social conventions and institutions both facilitate and depend on human capacities for social coordination.

2a) Social coordination and social cognition

Human beings live in group sizes that outscale any other animal society, apart from genetically closely related superorganisms, such as ants and bee hives (Wilson 2012). The average proximate group size (involving regular contact and name recognition) of most humans is on the scale of 150 individuals (Bzdok&Dunbar 2020, 720), thousands of individuals cross each other on a daily basis when commuting, and millions of individuals may occupy the same urban territory. Such large-scale social life demands coordination: in order to avoid hostile confrontations and accidental collisions, individuals need to be able to coordinate their attitudes and behaviors in a manner that is mutually beneficial, or minimally, avoids unnecessary harm. How do we achieve this remarkable feat?

The nature of the human capacity for social coordination, as well as the cognitive and social mechanisms underlying its manifestations in various environments, have become prominent topics of research in many different scientific disciplines in recent years. These disciplines include (but are not limited to) analytic philosophy (e.g. Bratman 2014; Gilbert 2013; Miller 2001; Searle 2010; Tuomela 2013), cognitive science (e.g. Dumas et al. 2021; Frith&Frith 2012; Gallotti et al. 2017; Knoblich et al. 2011; Michael et al. 2016; Vesper et al. 2016), developmental psychology (e.g. Carpenter&Svetlova 2016; Rakoczy 2017; Spelke et al. 2013), experimental economics and (evolutionary) game theory (e.g. Bowles&Gintis; Guala 2016; Ross 2011; 2012; Roth 2016; Schelling 1978), and evolutionary anthropology (e.g. Boyd&Richerson 2005; Dunbar 2009; Henrich&Muthukrishna 2021; Hrdy 2019; Tomasello et al. 2005). These disciplines engage with human social coordination on multiple spatial and temporal scales, from the phylogenetic time scale of gene-culture coevolutionary processes (e.g. Boyd&Richerson 2005; Henrich&Muthukrishna 2021) via the ontogenetic time scale of individual development (e.g. Carpenter&Svetlova 2016; Siposova&Carpenter 2019) to the synchronic time scale of two physically co-present individuals coordinating their motor behaviors with one another (e.g. Gallotti et al. 2017; Dumas&Fairhurst 2021).

There has so far been relatively limited systematic engagement between different disciplinary approaches to social coordination, despite their common interests, and the fact that researchers are clearly aware of related work that is taking place in other disciplines (see e.g. Guala 2016; Knoblich et al. 2011; Tomasello 2019; Tuomela 2013). Of course, discipline-centric institutional incentives may play some role in hindering more extensive interdisciplinary collaboration (Frodeman et al. 2017; Mäki 2016, cf. Nagatsu et al. 2020). However, my dissertation

is premised on the idea that one important obstacle to greater interdisciplinary exchange and integration is the absence of common methodological frameworks that help us to understand how different disciplinary approaches to social coordination relate to one another, what types of methodological continuities or differences there are between them, how their theoretical outputs can be integrated with one another, and what types of challenges may arise in this process (as well as what apparent challenges turn out to be merely illusory or easily resolved). These are among the questions that I engage with in my dissertation.

To lead into my subject matter, the notion of *social coordination* can be understood in the most general sense as encompassing all situations in which the actions of two or more agents are responsive to one another. In this broad sense, the movements of a predator and her prey (e.g. a lion hunting an antelope, who is trying to escape) can be described as involving social coordination. However, in a more restrictive sense, social coordination is typically understood as involving at least some type of common goal or cooperation between the participants to social coordination. For example, the cognitive scientists Natalie Sebanz, Harold Bekkering and Günther Knoblich describe a *joint action* as "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" (Sebanz et al. 2006, 70). In a subsequent paper, Knoblich and Sebanz together with the philosopher Stephen Butterfill distinguish between *emergent* and *planned coordination*. They say that "planned coordination… is driven by representations that specify the desired outcomes of joint action and the agent's own part in achieving these outcomes", while "in emergent coordination, coordinated behavior occurs due to perception–action couplings that make multiple individuals act in similar ways… independent of any joint plans or common knowledge" (Knoblich et al. 2011, 62). My focus in this

dissertation is on planned coordination, and especially forms of planned coordination that involve what philosophers have described as *shared intentionality* (more on this in the next section).

The capacity for planned coordination, and for joint action in particular, in most circumstances depends on the capacity for *meta-representation*—i.e. the capacity of individuals to attribute psychological states, such as beliefs, goals, and/or intentions to their co-actors (Bratman 2014; Butterfill 2020; Tuomela&Miller 1988; cf. Colombo&Guala 2021). However, there is considerable controversy among cognitive scientists about the precise nature of this capacity, its limits, and the psychological mechanisms underpinning it (Apperly&Butterfill 2009; Christensen&Michael 2016; Goldman 2008). Some cognitive scientists have argued that meta-representation involves appeal to *unobservable* posits (e.g. beliefs and desires) that are used to explain observable behavior in a manner that is analogous to how scientists appeal to unobservable forces (e.g. gravity) to explain observable phenomena, (e.g. the motions of the planets in the solar system) (Gopnik&Meltzoff 1997; Lewis 1972; Sellars 1963). Others have argued that much of our ordinary, folk psychological action understanding is achieved through a cognitively less demanding process of *mental simulation*, where one's own decision-making system is used as a mental model of how the other is likely to behave (Goldman 2008). Yet others have sought to straddle these views by arguing that our ordinary framework of intentional agency is best understood as a *family* of loosely connected (embodied and theoretical) models that are used flexibly in the pursuit of a broad range of practical and theoretical goals (Godfrey-Smith 2005; article II of my dissertation). In my dissertation, the analysis of people's meta-representational capacities in terms of their capacity to attribute different types of mental (or as philosophers often say: *intentional*) states to one another forms an important point of contact between studies of social coordination in the behavioral and cognitive sciences and

philosophy.[1] In the following two sections, I will discuss philosophical work on individual and shared intentionality in some more detail, before returning to the ontogenetic and phylogenetic development of human capacities for social coordination in section 2d.

2b) Social coordination and shared intentionality

The notion of *intentionality* has deep roots within philosophy, and is often traced back in its modern form at least as far as the work of the 19th century phenomenologist Franz Brentano (1874). However, the roots of this notion can even be identified in the work of Aristotle and the Medieval scholastics (Caston 1998; Crane 1998). The phenomenon of intentionality is sometimes glossed as 'aboutness', and can be elaborated in some more detail as "the capacity of agents to entertain contentful attitudes (beliefs, desires, intentions, etc.) towards the world and to be guided by these in reasoning and action" (Rakoczy 2017, 139). In the philosophy of action, different types of intentional attitude types are often distinguished by the functional roles that they play in the practical reasoning, deliberation, and behavior of agents (Braddon-Mitchell&Jackson 2007; article I of my dissertation). For example, a "desire that p" has been described as an intentional attitude that leads to behavior that is conducive to bringing about that p (provided that there are no stronger countervailing desires, and the agent knows how to bring about p), while a "belief that p" has been described as an intentional attitude that (among other things) leads to the inference that

[1] There is a long tradition of exchange between analytic philosophers and behavioral and cognitive scientists about the mechanisms underpinning our meta-representational capacities going back at least to Dennett's (1978) and Premack's&Woodruff's (1978) debate about whether chimpanzees have a theory of mind—and how to determine whether they do.

q, if the agent also believes "if p then q" (Bratman 1987; Harman 1986; Searle 1983). As the latter example indicates, intentional attitudes are individuated in part by the inferential roles that they play in the practical reasoning and deliberation of agents, not merely by their behavioral outcomes. For this reason, many philosophers have argued that there is always some degree of interpretation involved in the ascription of intentional attitudes (e.g. Davidson 1980).

The term *shared intentionality* is often used in analytic philosophy to refer to the attribution of goal-directed actions and/or mental states to collections of individuals (e.g. Bratman 1999; 2014; Gilbert 1990; 2013; Searle 1990; 2010; Tuomela 2007; 2013; cf. Sellars 1963). In this respect, *distributive* forms of shared intentionality (e.g. some members of Parliament believe that same-sex marriage is morally prohibited) have been distinguished from *non-distributive* forms of shared intentionality (e.g. the Parliament has passed a resolution to legalize same-sex marriage) (Ludwig 2016; 2017; cf. Quinton 1975-76). As stand-ins for the notion of shared intentionality, also the terms 'collective', joint', and 'we-intentionality' have been used. Although these terms have slightly different connotations for different authors (see e.g. Tomasello 2019), I will treat these terms as equivalent in my dissertation.[2] While some authors describe shared intentionality in a terminologically loaded manner—for example, as encompassing situations in which "two or more agents form a joint "we" attitude in a way that is not

[2] Many philosophers (e.g. Searle 1990; 2010; Tuomela 2007; 2013) and behavioral scientists (e.g. Gallotti&Frith 2013; Tomasello 2019) seem to take the terms 'collective intentionality' and 'we-intentionality' to involve a bias in favor of irreducibility, while the term 'shared intentionality' is terminologically more neutral or reductionistically oriented (e.g. Bratman 1999; 2014). The terms 'joint intentionality' and 'joint action have been more common among cognitive scientists and philosophers influenced by cognitive science (e.g. Knoblich et al. 2011; Sebanz et al. 2006; Vesper et al. 2016).

straightforwardly reducible to mere sums of individual intentional attitudes" (Rakoczy 2017, 141)—such expressions tend to be biased in favor of a non-distributive account of shared intentionality, and therefore will not do a as a general definition. My view is that shared intentionality is best understood as encompassing not only a broad range of different psychological attitudes, such as shared beliefs (e.g. Hakli 2006), shared emotions (e.g. Salmela 2013), and shared intentions (e.g. Bratman 2014), but also a broad range of different ways in which such attitudes can be shared (e.g. distributively or non-distributively).

The notion of shared intentionality can be used to ground a distinction between *shared intentional actions* (which cognitive scientists often refer to as joint actions—see previous section) and unintentionally coordinated behaviors of two or more individuals. As several philosophers have noted (e.g. Gilbert 1990; Bratman 2014), there seems to be an important intuitive difference between cases such as two friends going for a walk together and two strangers walking alongside one another on a busy street, even if the latter are coordinating their movements skillfully so as to avoid collisions. According to most philosophical accounts, this difference comes down to relevant *shared intentions* of the participants (Bratman 1999) and their *mutual beliefs or expectations* concerning each other's activity (Tuomela&Miller 1988; Bratman 1999)[3], where the latter

[3] For example, Bratman (2014, 10) writes: "Echoing Wittgenstein's question about the difference, in the individual case, between my arm's rising and my raising it, we can ask: what is the difference between such a contrast case and corresponding shared intentional activity? In the case of individual intentional human action, we can see the difference from a contrast case as involving an explanatory role of relevant intentions of the individual agent. As a first approximation, I propose an analogous view of the shared case: the difference in the case of shared agency involves an appropriate explanatory role of relevant *shared* intentions. Our painting together is a shared intentional activity, roughly, when we paint together because we share an intention so to act."

may also involve some *normative expectations* about the circumstances under which one it is acceptable to detach oneself from the joint activity in question (Gilbert 1990; 2013). However, there is considerable controversy about how to analyze shared intentions in more detail (see article II of my dissertation).

To identify a useful contrast class, it is helpful to consider some further forms of social coordination that do not depend on shared intentionality. For example, ants are known to coordinate many of their behaviors on the basis of pheromone trails, while bees coordinate on a new site for a nest on the basis of an elaborate sequence of dance-like moves (Seeley 1995). These forms of social coordination seem to be based on reflex-life responses to chemical signals or other sub-intentional cues, rather than on "the capacity of agents to entertain contentful attitudes… towards the world and to be guided by these in reasoning and action" (Rakoczy 2017, 139). Also humans sometimes coordinate their behaviors in something like this manner. For example, individuals have been shown to sub-consciously align their bodily postures, mannerisms, and facial expressions with one another (Chartrand&Bargh 1999; Lakin&Chartrand 2003; Richardson et al. 2007). In addition to such forms of subconscious interpersonal alignment, social coordination can sometimes be based on the deliberate scaffolding of the environment in a manner that is conducive to social coordination—for example, by designing crowded spaces so as to reduce collisions, or by designing market institutions that make buyers and sellers more likely to find one another and to converge on common price (Roth 2016). My focus in this dissertation is on forms of social coordination that do depend on shared intentionality, and therefore I will not discuss topics such as market design or the ecology of urban spaces.

2c) Philosophical approaches to shared intentionality

In accordance with the dominant approach to analytic philosophy during the 20th century (Soames 2003), many philosophical approaches to shared intentionality have built on the conceptual analysis of natural language sentences, such as "We intend to X" (e.g. Ludwig 2016; Tuomela&Miller 1988). Such sentences seem to involve at least three distinct parts, which can be identified on an analogy with individual-intentionality attributing sentences. Consider the sentence "John intends to walk to the fair". On the face of it, this sentence involves a particular agent ("John"), a particular attitude type ("intention"), and a particular action content ("walking to the fair"). Analogously, the sentence "We intend to walk to the fair together" seems to involve a particular collection of agents ("We"), their relevant attitudes (e.g. part-performance intentions and mutual beliefs), and a particular shared action content ("walking to the fair together"). In accordance with this tripartiate structure, many philosophers have sought to elucidate the types of agents, attitudes, and/ or action contents that are implied or presupposed by linguistic and conceptual attributions of shared intentionality.

Raimo Tuomela (2007; 2013) is one philosopher, who analyzes shared intentionality by appeal to the idea of a group agent. His account focuses on robust forms of shared intentionality in what he calls the *we-mode*, as contrasted with weaker, *I-mode* forms of shared intentionality. Tuomela (2013, 15-16) says that the *intentional agent* of shared intentionality in the we-mode is the group that the individuals jointly constitute. Tuomela (2013, 47-48) characterizes a we-mode group agent as a *partly fictitious* entity that is *collectively constructed* to serve as the representative of attitudes and actions that are attributable to the individuals *as a group*. According to Tuomela (2013, Ch.2), the members of a we-mode group agent are expected to take their *reasons for acting*

from what the group agent believes, desires, and intends, to be *collectively committed* to the group's activities, and to satisfy the *Collectivity Condition*, according to which the goals and concerns of the group agent are satisfied for each of the group members *if and only if* they are satisfied for each of the other members. For example, our we-mode intention to clean the park together might provide each group member with a reason to pick up litter, even if we have no private desire to do so (Tuomela 2013, 131). Still, all group members get to enjoy the satisfaction of the shared intention, including the benefits of strolling about in a clean park, regardless of whether they actively contribute to bringing about this goal or not.

Margaret Gilbert (2013) *plural subject* -account develops the idea of a group agent in a manner that contrasts with Tuomela's account. According to Gilbert (2013, 41), shared intentionality involves a *joint commitment* of several individuals to "emulate, by virtue of the actions of each… a single body that intends to do the thing in question". Gilbert (2013, 41) argues that a joint commitment comes into being *simultaneously* and *interdependently* for all of the participants to the joint commitment once they have each expressed, in conditions of common knowledge, their willingness to be jointly committed to a common goal. Gilbert believes that a joint commitment brings into existence *sui generis*, non-instrumental and non-moral rights and obligations between the participants to shared intentionality. For example, we might form a joint commitment to go for a walk as a result of a discussion about what to do this afternoon (Gilbert 1990). If either of us were to renege on our joint commitment, or turn back without notifying the other person, she would owe the other an explanation, and the other would have the standing to rebuke her.

John Searle (2002; 2010) rejects appeal to irreducible group agents on grounds of metaphysical parsimony. His account of *we-intentions*

capitalizes on the familiar distinction between the form (or "mode") and content of intentional states (Searle 1983). According to this distinction, the same propositional content (e.g. "that there is beer in the refrigerator) can be held in different psychological modes (e.g. 'belief' and 'desire'), which give rise to different action dispositions (e.g. the disposition to go to the refrigerator for a cold beer or to go to the grocery store to stock up on refreshments). Extending this distinction further, Searle (2002) argues that individuals may also have psychological states with an irreducibly collective form (e.g. 'we-intentions' or 'we-beliefs'). Searle (2002, 97) argues that the collective form of *we-intentions* can be represented outside the brackets that specify the propositional content of the intentional state, as expressed in the schematic form "We intend (X)". When such generic we-intentions are filled in with suitable collective action types replacing the X-term inside the brackets, appropriate part-performance intentions can be derived from them on the basis of contextually available information. In Searle's (2002, 100) example, I derive from my we-intention to prepare hollandaise sauce together with you the intention to do so by pouring, while you derive from your respective we-intention to prepare hollandaise sauce together with me the intention to do so by stirring.

Michael Bratman (1999; 2014) rejects appeal to irreducible group agents or sui generis we-intentions in his constructivist account of small-scale forms of shared intentionality in the absence of asymmetric authority relations, such as two people singing a duet together. Bratman argues that we can scale up from such basic forms of *modest sociality* to more complex and hierarchically structured forms of shared intentionality within institutional circumstances, such as an orchestra performing together under a conductor. According Bratman (1999, 121), it is sufficient for our shared intention to act together that each of us intends that we J *in accordance with* and *because of* each of our intentions that we J, and all of this occurs in conditions of common knowledge (where J

denotes a joint action type). Bratman (2014, 85-86) later builds upon this original account by adding a number of further conditions, which increase the stability and depth of shared intention. These additional conditions include, among other things, interdependence in the persistence of our intentions in favor of J-ing, and intentions on the part each of us that our sub-plans mesh. Bratman (2014, 11) characterizes his account of modest sociality as a form of *augmented individualism*, because it draws on special semantic contents and explanatory interconnections between the intentional states of several individual agents. However, by contrast to the accounts of Gilbert, Searle, and Tuomela, it does not appeal to *sui generis* types of intentional states ('we-intentions') or irreducible group agents.

In addition to the well-known accounts of shared intentionality that have been discussed above, a number of *minimalist* accounts of shared intentionality have been put forth (e.g. Butterfill 2012; Michael 2011; Pacherie 2013; Saint-Germier et al. 2021; Tollefsen 2005). Some of these accounts have challenged the assumption that shared intentional action requires appeal to shared intention, while others have proposed less demanding accounts of shared intention that can accommodate the shared intentional actions of cognitively relatively unsophisticated agents, such as young human infants (Butterfill&Sebanz 2011). Often, these accounts are cognitively less demanding in the sense that they do not require individuals to be able to formulate propositionally structured higher-order beliefs or intentions about one another's intentional states, a capacity that infants do not seem to fully possess until around four years of age (Carpenter&Svetlova 2016; Rakoczy 2017). While I will not discuss minimalist accounts of shared intentional action here in more detail, many of the ideas that I formulate in my dissertation can also be extended them (see esp. article IV). Moreover, I share with the minimalists the idea that our empirical understanding of the nature and mechanisms of human social coordination sets

constraints on what type of philosophical account of shared intentionality can be regarded as acceptable. Thus I will use the next section to discuss the ontogenetic and phylogenetic development of human capacities for social coordination in some more detail.

2d) Ontogeny and phylogeny of social coordination

In addition to analytic philosophy, there has been a social turn across the behavioral and cognitive sciences in recent decades, where the focus has gradually shifted from individual behavior and cognition to the types of behavioral outcomes and cognitive requirements that are characteristic of multiple individuals interacting in the same situation or in pursuit of a common goal (see e.g. Gallotti et al. 2017; Dumas&Fairhurst 2021; Knoblich et al. 2011; Michael&Pacherie 2015; Tomasello et al. 2005; Tomasello 2019, Vesper et al. 2016). While my focus in this dissertation is on philosophical studies of shared intentionality, there have been important two-way interactions between them and related research in disciplines such as cognitive science, developmental psychology, and evolutionary anthropology during recent years. On the one hand, philosophical analyses of "we-intentionality" (Searle 1990; Tuomela&Miller 1988; Tuomela 2007) have inspired research on a special "we-mode" of cognition in cognitive science (Gallotti&Frith 2013; cf. article III of my dissertation), as well as theories of the uniquely cooperative nature of human cultural intelligence in evolutionary anthropology (Tomasello et al. 2005; Tomasello 2019; see article V of my dissertation). On the other hand, empirical research on human social coordination has supported particular philosophical models of shared intentionality (see e.g. Butterfill&Sebanz 2011; Butterfill 2012; Godman 2013; Michael 2011; Tollefsen 2005; article IV of my dissertation). Thus it is reasonable to

spend some time discussing the ontogenetic and phylogenetic background of human capacities for social coordination in more detail.

To begin with a phylogenetic perspective, the idea that human individuals are biologically predisposed for mutually beneficial social interaction with other human individuals, especially with those whom they identify as members of their in-groups, is shared by a large group of evolutionary anthropologists today. For example, Robin Dunbar (1992; 2009) has provided evidence that brain size (and especially neocortex size) correlates with group size in a large number of primates, suggesting that a significant part of our brain capacity is used for processing information about social relations. Michael Tomasello (Tomasello et al. 2005; Tomasello 2019) has proposed that human beings posess species-distinctive skills and motivations for shared intentionality, which he describes as "collaborative interactions in which participants have a shared goal (shared commitment) and coordinated action roles for pursuing that shared goal" (Tomasello et al. 2005, 680). And Brian Hare (2017) and Richard Wrangham (2019) have argued that human beings are characterized by reduced reactive aggression (compared to most of our nearest primate relatives) and a type of generalized docility towards other in-group members, including genetically unrelated individuals. These capacities appear to be at least in part distinctively human in the sense that our nearest primate relatives, the chimpanzees, are much more competitively oriented in their interactions than humans (Hamann et al. 2011; Hare&Tomasello 2004; for an apparent exception concerning the bonobos, see Wrangham (2019)).

The precise nature of the ecological environments in which human capacities for social coordination evolved is hard to pinpoint with certainty, because of the scarcity of evidence regarding our evolutionary past. However, it seems plausible that they may have first evolved in

the context of collaborative foraging and hunting practices, which were in turn driven by some change in the environment, which forced human groups to search for new sources of food or to compete more fiercely for existing sources of food with other species and/or other hominid groups (Tomasello 2019, 5). Such collaborative hunting practices may in turn have required changes in the social structure of groups, such as cooperative breeding practices involving fathers, grandparents, and siblings as caretakers in addition to mothers, as well as a more advanced division of labor that is sensitive to age and gender differences (Hrdy 2009). They may also have required individuals to be willing to share the spoils of collaboration and to be more tolerant of genetically unrelated individuals in general (Hare 2017; Wrangham 2019). The evolution of these distinctive capacities for social coordination is likely to have been driven at least in part by gene-culture coevolutionary processes (Boyd&Richerson 2005), where more frequent and intense social interactions drove the need for new genetic adaptations that were suited to the changing social environments that humans had created (Henrich&Muthukrishna 2021):

> "By generating increasingly complex tools (e.g. spear- throwers), food processing techniques (e.g., cooking), languages (e.g., larger vocabularies) and institutions (e.g., clans) over hundreds of thousands of years, cumulative cultural evolution has shaped the environments faced by our genes and thereby drove the genetic evolution of the uniquely human aspects of our bodies and minds. For example, our small stomachs, short colons and weak jaw muscles, compared to other primates, were only favored once fire and cooking had spread culturally in our species…"

To move on to an ontogenetic perspective, the most basic forms of social coordination that human infants engage in from within a few days of birth have been described as forms of *primary intersubjectivity* (Trevarthen&Aitken 2001). While already neonates are sensitive to the

body temperature of their mother and the smell of milk, the most basic forms of human social interaction that take place immediately after childbirth (e.g. in the form of sucking behavior) are likely to be mostly hard-wired and reflexive. Within a few weeks of birth, human infants seem to acquire some degree of flexibility in their social interactions with others when they begin to search and maintain eye contact and to imitate their caretakers' facial gestures, for example, by smiling or protruding their tongues when their caretaker produces a similar gesture (Meltzoff&Moore 1977; cf. Davis et al. 2021). Gradually, this phenomenon of *neonatal imitation* gives way to temporally more structed interactions in the form of *protoconversations* (Bateson 1973), where the infant and her caretaker take turns producing relevant facial or gestural expressions and imitating one another. Stephen Levinson (2016) has argued that the type of non-linguistic turn-taking that is manifest in protoconversations is a cultural universal, which serves as one important foundation to the development of natural language. The importance of social coordination during the first months of an infant's development (Trevarthen&Aitken 2001) is underscored by behavioral and hormonal signs of distress during *still-face experiments*, where a caretaker abruptly stops responding to the behaviors of the infant (Tronick et al. 1978).

The next important step during an infant's social development takes place during the *nine-month revolution* (Tomasello 2019, 55-56), when infants begin to engage in joint attention and to engage in simple social games with their caretakers. *Joint attention* can be understood as the coordination of attention with others towards an object of mutual interest (Bakeman&Adamson 1984), as exhibited, for example, in the form of pointing behavior or common awareness of something interesting that is happening in the environment (imagine watching a wild deer that is walking across the yard). Thus joint attention is not simply attending to the same thing, either deliberately (as in gaze

following) or non-deliberately (as when a loud noise startles everyone), because in these phenomena intentional coordination is lacking (Siposova&Carpenter 2019). *Social games* are simple collaborative activities, which serve no further instrumental goal apart from the establishment or maintenance of a social bond and the intrinsic pleasure of the joint activity itself, for example, when bouncing a ball together with a partner on a large trampoline (Warneken et al. 2006). Interestingly, while Warneken et al. (2006) found that chimpanzees were able to coordinate instrumentally with human partners in a task involving relatively sophisticated division of labor in order to obtain a reward (an item of food), they were unable to engage chimpanzees in much more simple social games at all (see also Tomasello&Carpenter 2005). This suggests that even if non-human primates have the same basic cognitive capacities for social coordination as we do, they lack the types of pro-social motivations, which make these capacities become manifested in particular ways (Godman 2013; article V of my dissertation).

The third crucial stage in the development of human capacities for social coordination takes place beginning from the third year of an infant's life. During these years of rapid social development, the types of triadic interactions (involving you, me, and a shared object of our attention) that first emerge during the nine-month revolution become increasingly more complex due to an increasing understanding of the types of social norms that are characteristic of acting together (see Michael et al. 2016). For example, Gräfenhain et al. (2009) found that 3-year old infants attempt to re-engage their collaborative partners and excuse themselves when quitting a joint activity, but only when they have jointly committed (either verbally or non-verbally) to engaging in the activity together. Hamann et al. (2011) found that collaboration induces equal sharing of rewards in 3-year old human infants, but not in chimpanzees. Kachel et al. (2019) found that 3-year old infants react

differently to the sudden interruption of a joint activity when it is due to intentional defection, ignorance, or an external disturbance, protesting normatively only when appropriate. And Lyons et al. (2007) showed that human infants tend to overimitate the activities of others even when this is non-conducive to or maladaptive relative to the instrumental goal of the activity—e.g. performing an extraneous physical movement prior to performing an instrumental task—suggesting that they do not care only about the instrumental goals of the activity, but also about what is the normatively appropriate way to act.

The normative scaffolding of joint activities in early infancy finds resonance in philosophical approaches to shared intentionality that have emphasized the importance of joint commitments (Gilbert 1990; 2013) or group-social normativity (Tuomela 2013) as central determinants of social coordination (see also Gomez-Lavin&Rachar 2019; 2021). However, it also has a connection to more explicitly functionalist and plan-theoretic accounts of shared intentionality, which emphasize the roles of shared intentions in structuring relevant shared deliberation and planning (e.g. Bratman 2014). For example, the cognitive scientists John Michael, Natalie Sebanz and Günther Knoblich argue that "commitments make individuals 'behavior predictable in the face of fluctuations in their desires and interests, thereby facilitating the planning and coordination of joint actions involving multiple agents… commitment also facilitates cooperation by making individuals willing to contribute to joint actions to which they wouldn't be willing to contribute if they, and others, were not committed to doing so" (Michael et al. 2016, 1). Nevertheless, philosophical debates about whether shared intentionality *essentially* involves special forms of group-social normativity that go beyond instrumental means-ends normativity (see e.g. Bratman 2014; Gilbert 2013) strike me as somewhat inconsequential. Even if our ordinary, folk

psychological understanding of joint commitment is associated with a sense of mutual entitlements and obligations, it might still be justified to abstract away from such entitlements and obligations for the purposes of modeling certain basic forms of shared intentionality. This is an upshot of the model-based approach to shared intentionality that I defend in several articles in my dissertation (esp. articles II and III).

2e) From social coordination to social ontology

Social conventions and institutions facilitate social coordination in numerous ways (Aoki 2001; Greif 2006; Guala 2016; Miller 2010; North 1990; Tuomela 2013). For example, without the existence of money as a generalized medium of exchange, store of value, and unit of account, coordinated division of labor on a society-wide scale would be extremely difficult, since individuals would be largely limited to barter or cross-temporal exchange with a limited group of trusted individuals (Menger 1982; Mäki 2020b; Smit et al. 2011). Similarly, a common language qua a complex cultural institution facilitates bargaining and negotiation over coordinated action plans (Bratman 2014), and makes possible the formulation of written agreements that codify and keep a tab on the respective responsibilities of participants (Bach 1995). The evolution of language may in turn be viewed as a complex coordination problem in its own right, where the challenge is for a community of language-users to coordinate on a common set of symbols for representing phenomena that are in some sense accessible to all (Lewis 1969; Skyrms 2010). Social institutions also facilitate collectively beneficial solutions to collective action problems, such as the Tragedy of the Commons and the Prisoner's Dilemma, through social norms that are associated with internal (e.g. feelings of guilt) or external (e.g. fines or other criminal penalties) sanctions that motivate individuals to conform to the norms in question (Bicchieri 2006; Hardin 1968;

Ostrom 1990). Moreover, strong public institutions have been argued to be responsible for significant differences in economic growth between countries (Acemoglu&Robinson 2012). Thus well-functioning institutions arguably promote human welfare in many ways (Hindriks&Guala 2019; Miller 2010).

The nature and functioning of social institutions, understood broadly as norm-governed social practices, such as marriage or private property, has been one of the central concerns of social ontology, a branch of philosophical investigation that is concerned with "the study of the nature and properties of the social world… concerned with analyzing the various entities in the world that arise from social interaction" (Epstein 2018, 1). According to an influential view in social ontology, social institutions must be *collectively accepted* (Searle 1995; Tuomela 2007; 2013)—or in the case of dysfunctional or unfair institutions, such as apartheid, *collectively recognized* (Searle 2010)—by the majority (or at least an influential minority) of the participants to the social practices that the institution governs.[4] Collective acceptance confers deontic rights and obligations on the participants to these practices—for example,

[4] Collective acceptance -based accounts have often formalized institutional properties in terms of constitutive rules with the form "X counts as Y in C" (Searle 1995)—where X picks out a non-institutional property (e.g. a river that separates two territories), Y picks out an institutional property that confers normative rights and obligations on some individuals (e.g. a border that must not be crossed by those who live in one territory), and C picks out relevant circumstances in which the normative rights and obligations in question are exercised (e.g. in times of peace or provided that relevant travel documents have (not) been accorded). Frank Hindriks (2009) has argued that constitutive rules can be analyzed in terms of collections of regulative rules of the form "If C, do X". Guala&Hindriks (2015; see also Hindriks&Guala 2015) have put forth an account of social institutions as *rules-in-equilibrium*, which aims to combine the virtues of collective acceptance - and equilibrium -based accounts of social institutions (see also Greif&Kingston 2011).

numerous formal and informal norms govern the institution of marriage, including informal norms against adultery and the duty (that is enshrined in law) to care for one's spouse financially. Some philosophers (e.g. Searle 1995; Tuomela 2007) have even argued that collective acceptance is required for all social institutions.[5] My own view is that collective acceptance (or recognition) may well explain some aspects of social institutions—notably, the proximate psychological mechanisms underlying their persistence or stability—but they need to be accompanied by other approaches in order to explain the emergence of social institutions and institutional change. As Aydinonat&Ylikoski (2018) point out, there are many features for a theory of social institutions to explain, and it may be unrealistic to expect one grand theory to explain all of these features.

An alternative to collective acceptance -based accounts of social institutions has been provided by game-theoretic accounts of equilibrium behavior (see Guala 2016; Lewis 1969; Leyton-Brown&Shoham 2008). When a profile of actions is in *equilibrium*, each player's action is the best response to the action of the other player, such that no player has an incentive to deviate from the profile of actions that make up the equilibrium. Social conventions, such as driving on the left or right hand side of the road, form an important class of game-theoretic equilibria, where the optimal choice of each player depends on the choices of the other players. In this case, each player prefers to drive on the right or left hand side of the road on condition that each of the other players drives on the same side of the road, but the players are indifferent about which side of the road they drive on (Guala 2016; Lewis 1969). When faced with the problem of equilibrium selection (and communication is not an option) game

[5] Tuomela (2007, 183) writes that "we-mode collective acceptance creates, and is required for, institutional entities and practices".

theorists often assume that individuals can coordinate on one of several alternative equilibria by recognizing some equilibrium as salient, where the salience of an alternative can be based on the history of play (Lewis 1969), some external signal (Gintis 2009; Guala 2016; Vanderschaaf 1995), or common background knowledge (Schelling 1960). While coordination in the driving game seems relatively straightforward (at least in a simplified normal-form, two-person version of the game) additional complications are brought in when the interests of the players are imperfectly aligned, as in the infamous Battle of the Sexes, where a husband and a wife attempt to resolve whether to watch a romantic comedy or an action movie, or in the Hawk-Dove, which has been used by evolutionary game theorists to model the evolution of the institution of private property (Gintis 2007; Maynard-Smith 1982).

My approach to social coordination in this dissertation does not rely explicitly on game theory, but I have been influenced by game theoretic models of social coordination in other ways. In particular, I have been impressed by how many abstractions and idealizations game-theoretic models of social coordination involve relative to the social cognition and behavior of actual agents—whether it be the assumption of perfect strategic rationality (Amadae 2015) or common knowledge assumptions (Aumann 1995)—and the types of parallels that these have with philosophical research. Indeed, although many philosophers ground their accounts of shared intentionality in "folk psychological" terms and concepts (e.g. 'belief' and 'intention' instead of 'credence' or 'utility function')— also their accounts seem to embody many abstractions and idealizations that can be justified by considerations of simplicity or tractability that are comparable to the justification of unrealistic assumptions in game theory. However, while game theorists are typically aware of the pragmatic and stipulative nature of their assumptions, philosophers often think of themselves as revealing the 'essential' nature of 'paradigmatic' social phenomena, such as *we-*

intentionality (e.g. Tuomela 2013), *joint commitment* (Gilbert 2013), or *constitutive rules* (Searle 1995) through their conceptual analyses. If there are indeed important continuities between game-theoretic models of social coordination and philosophical work on shared intentionality—as numerous authors have argued during recent years (e.g. Bardsley 2007; Gold&Sugden 2007; Hakli et al. 2011; Pacherie 2013)—this tension needs to be resolved (see articles II-III in my dissertation).

3. Methodological approach of this dissertation

This part of my introductory essay discusses the general methodological principles that motivate my study of shared intentionality and social coordination. To begin, I will discuss the idea of interdisciplinarity as a constraint on theory formation in science, as well as the increasing focus on interdisciplinary during recent decades in science policy as well as in the philosophy of science. Then I will move on to theoretical modeling and model-construction as a distinctive approaches that scientists use to study complex phenomena, by constructing simplified surrogate systems that are used to stand in for the real system in question. Finally, I will discuss the framework of mechanistic explanation, which can be used to integrate explanations of related phenomena in different scientific disciplines that take place, metaphorically speaking, at different "levels of explanation" (e.g. behavioral, cognitive, ecological, and neural explanations of social coordination).

3a) Interdisciplinarity and philosophy of science

The idea of interdisciplinarity can be described as a productive constraint on theory formation in science, which has grown increasingly prominent during previous decades against the background of

increasing fragmentation of science into separate disciplines during much of the 19th of 20th centuries (Apostel 1972; Frodeman et al. 2017; Macleod 2018). While calls for interdisciplinarity are often associated with casting disciplines in a parochial light, there is an air of paradox in the claim that interdisciplinarity is needed in order to break disciplinary "silos" of knowledge, because calls for interdisciplinarity have also given rise to new research programs with a narrower focus than traditional disciplines, such as sustainability science (Nagatsu et al. 2020) or indigenous studies (Koskinen&Rolin 2019). The type of top-down, managerial interdisciplinarity that has given rise to new scholarly programs, as well as the types of institutional relations that are reflected in the allocation of power and resources between academic departments, play only a small role in the articles in my dissertation. However, the integration of knowledge from different scientific disciplines is a central underlying theme in many of the cases that I study. This section will discuss the types of challenges and opportunities that are related to interdisciplinarity in general terms, before considering some shared methodological frameworks that can play a role in interdisciplinary exchange and integration.

The need for interdisciplinarity is often motivated by the existence of problems and challenges that are too complex for any single discipline to solve, such as climate change or demographic transitions in developing countries (Mäki 2016). However, interdisciplinarity need not arise as a result of practical problems that call for scientifically informed interventions on the natural or social world, but may simply arise as a result of several disciplines studying closely related phenomena in their own disciplinary research programs. Such partly overlapping (in terms of subject matter) but separate (in terms of the social organization of science) research programs may spontaneously give rise to questions about how they relate to one another, whether the insights that they provide are mutually compatible, and whether they could or should be

integrated in order to provide a more comprehensive understanding of the phenomena under investigation. While wholesale interdisciplinary integration need not always be either desirable or necessary, parallel investigations of related phenomena in different scientific disciplines do seem to pose constraints on the types of explanations that can be regarded as acceptable in any single disciplinary research program—consider e.g. microbiological and population-level studies of a disease epidemic (Broadbent 2013). Parallel studies of similar phenomena may also prompt new empirical hypotheses and research questions to arise in related disciplines, generating forms of interdisciplinary exchange that fall short of interdisciplinary integration in a more systematic sense.

To consider some common pitfalls related to interdisciplinarity, Frodeman (2017) has remarked that the idea of interdisciplinary is sometimes treated as little more than an empty honorific, which is intended to underscore the innovative, multifaceted or transgressive nature of the research that is being carried out. However, genuine interdisciplinarity should be distinguished from *multidisciplinarity*, which involves mere juxtaposition of different disciplinary approaches without the aim of integrating knowledge (Apostel 1972, 25; Klein 2017, 22-24). In addition, interdisciplinarity is often distinguished from *transdisciplinarity*, which involves the transgression of academic boundaries through the engagement of extra-academic partners, such as corporations or citizen scientists, in addition to academic researchers (Apostel 1972, 26; Klein, 29-30; Koskinen&Rolin 2019). Finally, interdisciplinarity itself can take various different forms, from the borrowing of concepts and ideas to shared methodological frameworks or active collaboration in the pursuit of practical or theoretical goals (Klein 2017, 22). While some of these goals have the aim of integrating knowledge, interdisciplinarity cannot be equated with integration, because coordination between functionally autonomous teams of researchers may arguably sometimes be a more effective strategy than trying to bring their contributions

together under a common framework. As an example of the benefits of relative disciplinary autonomy, Jacobs (2017, 37) mentions the ebola epidemic in Africa, where teams of epidemiologists, health care workers, microbiologists, and social anthropologists worked towards a common goal, but largely separately, to come to terms with the biological (e.g. disease DNA), cultural (e.g. mistrust of authorities), and public health (e.g. availability of qualified nurses or hospital beds) determinants of disease spread.

The increasing focus on interdisciplinarity during recent decades has also been reflected in the philosophy of science. Uskali Mäki (2016, 335) maps out a trajectory in the philosophy of science from general philosophy of science in the 1950s (applying equally to all scientific disciplines) to philosophies of the special sciences from the beginning of the 1980s (addressing issues that are specific to disciplines such as biology, cognitive science, or economics) and, since the turn of the millennium, to a philosophy of the relations between the sciences—i.e. of interdisciplinarity. The topics that I address in my dissertation lie somewhere between the second and third fields of inquiry identified by Mäki (2016), given that I proceed by studying methods and forms of inquiry that are germane to particular disciplinary research programs, and use this understanding to ask questions about their relations to other disciplines and research fields. An important challenge in this task is to not essentialize disciplines, which are themselves historically contingent and institutional entities that have changed their shape across time. For example, in Early Modernity, science was divided into the fields of natural philosophy (including chemistry, physics, and the like), natural history (including fields, such as botany and zoology), and moral philosophy (encompassing topics, such as economics and political science, in addition to ethics). While the 19th and 20th centuries were periods of increasing disciplinary separation, the turn of the millennium has given rise to

both attempts to transgress disciplinary boundaries and to new interdisciplinary fields of study.

The most straightforward way to distinguish scientific disciplines from one another without overlooking historical differences in disciplinary boundaries is in partly institutional terms as "forms of social organization that evaluate, organize, and disseminate research and scholarship" (Jacobs 2017, 26). The conventional nature of this institutional criterion indicates that not too much weight should be put on disciplinary context. What is of most interest in interdisciplinarity from a philosophy of science perspective is relations between separate bodies of knowledge that are concerned with the same phenomena, whether they come from different sub-fields of one discipline or from separate disciplines altogether. In my dissertation, the shared epistemic goals of different disciplinary research programs are central to bringing together their contrasting perspectives on social coordination. While some of them may approach their subject matter in different ways or study different aspects of the same phenomena, the achievement of their shared epistemic goals can be facilitated by the use of common methodological frameworks, which facilitate interdisciplinary communication, exchange, and integration. Many of the articles in my dissertation are concerned with the articulation and elaboration of such frameworks (see especially articles I, III, V, and VI).

3b) Scientific modeling and model-construction

The relation between scientific models and theories has become a prominent topic of research in the philosophy of science during recent decades (see e.g. Craver 2002; Godfrey-Smith 2006a; Mäki 2009; Suarez 2004; Sugden 2000; Weisberg 2007a; 2013). While there are many different accounts of models and theories in the philosophy of science,

Peter Godfrey-Smith (2005, 2) has distilled currently widespread ideas about the relation between models and theories to three commonly shared sets of assumptions. First, talk of scientific theories in the philosophy of science has often been associated with the idea that there is a single way that all theorizing (in a broad sense) works within science, while discussions of scientific models have been associated with a more pluralistic understanding of the structure and methods of scientific investigation. Second, scientific theories have often been associated with the search for laws or universal generalizations, while models are associated with a looser relation of similarity (Giere 1988) or isomorphism (van Fraassen 1987) between the model and its real-world target. Third, scientific theories have often been taken to have truth-conditions and to make empirical claims about what the world is like, while scientific models make claims about the world only indirectly, by the mediation of *theoretical hypotheses* (Giere 1988) or *ontological construals* (Godfrey-Smith 2005), which specify in what respects and to what extent the model can be used to truthfully represent the world. In this section, I will discuss these three aspects of the relation between scientific models and theories in some more detail, before defending a particular view of scientific modeling as *indirect* investigation of the world by the mediation of a surrogate system, instead of studying the world directly, by way of observation or experiment (Godfrey-Smith 2006a; Weisberg 2007a; 2013).

The first point that Godfrey-Smith makes relates to the unity of science and the uniformity of scientific method. While it became popular during early Modernity to equate the scientific method with experimentation, contemporary philosophers recognize that there the process of science involves much more than empirical or experimental investigation. Some distinctive and quite heterogeneous activities that take place in scientific contexts include gathering data and representing it statistically (Suppes 1966), constructing concrete scale models or hypothetical structures that represent certain parts or aspects of the world (Downes

2011), and proving theorems or exploring mathematical structures (da Costa&French 1990). While the major achievements of science during the 19th and 20th centuries (e.g. Darwin's theory of natural selection, the atomic theory of matter, and quantum theory) gave rise to the idea that theories are the primary nexus of the scientific worldview, contemporary philosophers of science recognize that scientific understanding of the world may be conveyed by numerous different representational vehicles, from visual diagrams representing the building blocks of a cell (Bechtel&Richardson 2010) to experimental devices, such as the microscope (Hacking 1983), and other material artefacts such as maps or scale models (Downes 2011). Finally, while it is indisputable that our means of finding out about the world and representing it are heterogenous, some philosophers have gone even further than this and suggested that the world itself is ontologically disunified (Dupré 1993) or "dappled" (Cartwright 1999), and acquires a degree of lawfulness only as a result of our deliberate efforts to constrain reality into a uniform mold. Although I am not committed to this stronger thesis about the ontological disunity of the world, the heterogeneous nature of scientific methodology does play an important role in my dissertation (esp. article III).

The second point that Godfrey-Smith makes relates to the structure of scientific theories. The *received view of scientific theories* during much of the 20th century understood scientific theories as sets of universally quantified axiomatic statements, which were connected to empirical reality by bridge laws connecting the logical terms of the theory to observation sentences that were couched in descriptive vocabulary (Hempel 1965; Nagel 1961). The received view of scientific theories was associated with the deductive-nomological approach to scientific explanation, which understood explanation as consisting of the derivation of an observation statement from a statement picking out a general law and auxiliary conditions describing how the law manifests itself in particular circumstances (Hempel 1965). This approach was regarded as too

language-centric by many proponents of the *semantic approach to scientific theories* (Craver 2002; Suppe 1989), who identified a scientific theory with a family of models, which were described in set-theoretic terms (Suppes 1960), as trajectories in a state space (van Fraassen 1987), or in some other suitable mathematical vocabulary (Giere, 1988). The relation between such models and empirical reality was framed in terms of a relation of structural isomorphism (van Fraassen 1987) or the looser relation of similarity (Giere 1988) between a model and the world. While use of the term 'model' in the semantic view of scientific theories derived primarily from model theory in logic and its extensions (da Costa&French 1990), recent naturalist approaches to the philosophy of science have also sought to accommodate the looser talk of models that is common in many other disciplines across the behavioral and cognitive sciences (Giere 1988; Godfrey-Smith 2006a; Hausman 1992; Mäki 2009; Weisberg 2007a; 2013).

The third point that Godfrey-Smith (2005) makes relates to the representational relationship between scientific models and the world. While many early proponents of the semantic approach to scientific theories were happy to describe scientific representation as a two-place relationship between a model and the world, recent pragmatic approaches to scientific representation have emphasized that modeling should instead be understood as a three-place relation between the user of a model, the model itself, and the world (e.g. Giere 2004; Mäki 2009). The interests and purposes of the modeler have been invoked in order to overcome challenges that are associated with the looseness of the criterion of similarity (Giere 1988), since almost anything seems to be similar to almost anything in innumerable respects (Goodman 1976), and the too demanding nature of the requirement of isomorphism, because few things seem to be strictly isomorphic to one another (da Costa&French 2003). The pragmatic turn in philosophy of science brings questions about the nature of intentional agency to the forefront of questions

about scientific representation, forging a new link between the investigations of shared intentionality that this dissertation is concerned with and the philosophy of science.

The manner in which I have distilled the contrast between scientific models and theories above is conducive to a particular view of scientific modeling, which views scientific models as forms of *mediated representation* (Godfrey-Smith 2006a; Weisberg 2007a; 2013). Instead of directly studying the world through what Weisberg (2007a) has called *abstract direct representation*, scientific models represent the world only indirectly, through the mediation of *theoretical hypotheses* (Giere 1988; 1999; 2004), which specify in what respects and to what extent the model can be used to accurately represent the phenomenon under investigation. Thus scientific models allow for surrogate reasoning (Suarez 2004)—drawing inferences about the world by means of studying the model, rather than the world itself. Moreover, scientific models are able to accommodate many abstractions and idealizations, which may misrepresent some aspect of a phenomenon in order to bring other features into sharper relief (Mäki 2020; Thomson-Jones 2005; Weisberg 2007b; Wimsatt 2007). Finally, scientific models can form an object of study in their own right, independently of any real world targets, as models of perfect markets with omniscient agents or evolutionary dynamics with three-sex mating indicate (see e.g. Weisberg 2013). Given the mediated character of model-based representation, scientific modeling can be understood as a distinctive type of epistemic activity, which is not straightforwardly reducible to other forms of scientific theorizing (Godfrey-Smith 2006a; Weisberg 2007).

3c) Mechanisms and mechanistic explanation

The mechanistic approach to explanation in contemporary philosophy of science is based on the idea that many generalizations in science can be more adequately explained by describing underlying entities (or parts) and activities (as well as their interactions) that produce or underlie the phenomena that the generalization ranges over (Bechtel&Abrahamsen 2005; Craver&Darden 2013; Glennan 2017; Machamer et al. 2000) The central idea behind mechanistic explanation is accordingly to explain *why* some phenomenon occurs by describing *how* it occurs, or in the words of one classic paper, by describing "a structure performing a function by virtue of its component parts and component operations and their organization" (Bechtel and Abrahamsen 2005, 423). This account of mechanisms presupposes that mechanisms are at least weakly modular in the sense that they are decomposable into *structurally* distinguishable parts and *functionally* distinguishable activities that operate (at least in the short run) relatively independently from one another (Bechtel and Richardson 1993; Simon 1969). The mechanistic approach to explanation also naturally leads to the idea that mechanisms are hierarchically organized (or "nested") in the sense that the parts and activities that are responsible for a phenomenon can often themselves be further decomposed into constituent entities and activities, which in turn account for these lower-level phenomena (with the possible exception of fundamental physics).

Take, for example, the behavior of the human heart. The heart is part of the circulatory system, whose function it is to pump oxygen-rich blood and nutrients to the rest of the human body. The structural parts of the heart include (among other things) the four chambers of the heart, the valves which block reverse movement between the chambers, the vena cava through which oxygen-depleted blood enters the heart, the pulmonary artery through which blood is pumped to the lungs to be oxygenated, and the aorta through which the blood is ultimately pumped out of the heart into the arteries and on to the rest of the body.

There are several features of this *structural* decomposition of the heart that would not make sense without an appropriate *functional* decomposition of the heart into appropriate component operations: for example, the detour that the blood makes through the lungs, or the blocking of reverse movement by the valves, would seem functionally extraneous, if the function of the heart were to function as a boombox for making thumping noises, rather than to pump oxygen-rich blood and nutrients to the rest of the body. This has been taken by some philosophers to imply that mechanistic explanations are (at least) *epistemically non-reductive* in the sense that the parts and activities that constitute a mechanism are individuated in part relative to an appropriate functional description of the behavior of the system as a whole (Bechtel&Abrahamsen 2005).

The idea of mechanistic explanation may seem somewhat trivial in the context of phenomena that are well-understood by the standards of contemporary science, such as the functioning of the heart as part of the circulatory system. However, as a matter of historical fact, learning to understand the mechanisms of blood flow through the circulatory system was of course a substantial theoretical achievement, which required extensive empirical investigation and experimental ingenuity (Craver&Darden 2013, Ch. 7). Medieval anatomists following Galen believed that blood is continuously produced by the liver to be consumed by hungry tissue at the extremities of the human body. William Harvey challenged this received wisdom during the early days of the scientific revolution by conducting a series of experiments, ranging from dissections of live animals to ligatures used to constrain the flow of blood in the veins. Through these experiments, he identified central parts and operations that make up the circulatory system, and learned to understand the spatial, temporal, and functional constraints that govern its behavior.

The mechanistic explanation of complex behavioral traits may seem to involve challenges that go beyond the explanation of temporally and spatially relatively modular phenomena, such as the structure and functions of the human circulatory system. Consider the phenomena of memory (see Bechtel 2008; Craver 2010), aggression (see Longino 2013), or cooperation (see Bowles&Gintis 2011). One central challenge in explaining these types of complex behavioral traits has to do with appropriately individuating the phenomenon to be explained. For example, much psychological research on memory during the 20th century had to do with distinguishing between different memory systems, such as short-term and long-term memory, episodic and semantic memory, and procedural and declarative memory. Subsequent cognitive scientific research on memory has taught us about the types of functional constraints that are associated with different types of remembering (e.g. how many discrete chunks of information can be maintained in working memory (Miller 1956)), and about the brain mechanisms that are active in different types of memory systems (e.g. that long-term potentiation underlies many different types of remembering on the neuronal level or that the hippocampus plays a central role in spatial memory (Craver 2003)).

The mechanistic explanation of complex behavioral traits often requires interdisciplinary (or interfield) collaboration. Accordingly, Craver and Darden (2013, 12) argue that "the search for mechanisms provides a scaffold around which the findings of different scientific fields are integrated in a common explanatory objective". They distinguish between several different forms of interdisciplinary integration, from *simple mechanistic integration*, where "two or more fields investigate different stages in a mechanism, or different entities in a mechanism, or different aspects of the mechanism's organization" (Craver&Darden 2013, 163) to *intertemporal integration* and *interlevel integration*, which "links part to whole through a nested hierarchy of mechanisms within mechanisms" (ibid.).

Major scientific breakthroughs have often involved all three forms of interfield integration, as in the case of the *evolutionary synthesis*, which connected Darwinian mechanisms of natural selection across multiple generations of individual organisms to short-term molecular processes at the level of the gene (Craver and Darden 2013, 177-182).

The processs of mechanism discovery is typically a piecemeal and gradual endeavor, which admits not only of a division of labor between different scientific disciplines or groups of researchers, but also of different temporal stages. Often, the first stage in mechanism discovery is *phenomenal decomposition*, or identifying the phenomenon to be explained in a manner that is fine-grained enough for the purposes at hand, and distinguishing it from related but distinct phenomena. This process is far from trivial, as can be illustrated by psychological research on learning and memory during the 20th century, where distinctions between ever more complex types of learning processes gradually came to replace the behaviorist paradigm that was prevalent during the first half of the 20th century (Greenwood 2015). The second stage is *mechanistic decomposition*, or identifying the component entities and operations that give rise to the phenomenon in question. This process can be further distinguished into *structural decomposition*, or identifying component entities (e.g. the hippocampus as a central locus of spatial memory), and *functional decomposition*, or identifying component operations (e.g. the process of long-term potentiation on the neuronal level (Craver 2003)).

As a whole, the process of mechanism discovery may be viewed from an epistemic or an ontological perspective. From an epistemic perspective, a mechanism model can be described as *epistemically robust,* if it has been supported by findings from many different scientific fields, employing several different (types of) data, tractability assumptions, and/or research methods (Levins 1966; Kuorikoski et al. 2010; Weisberg 2006). Epistemic robustness typically supports *scientific realism*

about the ontological reality of the entities and activities that are represented by the mechanism model (Eronen 2015; Wimsatt 2007), and the assumption that the behavior of the system is *causally robust* in the sense that its behavior is relatively insensitive to changes in background conditions (Woodward 2006). To the extent that the phenomena represented by the mechanism model are causally robust and allow for non-trivial projections across a multitude of different circumstances, they may be described as picking out *natural* or *social kinds*, which can be expected to play an important role in the classificatory practices of science (Boyd 1991; Kuorikoski&Pöyhönen 2012; Godman 2020). From the perspective of mechanistic philosophy of science, we may accordingly ask whether studies of social coordination in different disciplinary research programs have been successful in identifying real and robust phenomena, which allow for novel projections, and constitute relevant kinds for the purposes of scientific investigation (see article V in my dissertation).

4. Main substantive claims of this dissertation

This part of my introductory essay motivates some of the main claims that I advance in my dissertation. First, I defend methodological naturalism as a general outlook on philosophical practice, which emphasizes important continuities between the methods of philosophy and scientific investigation, while allowing that such methods may form a rather heterogeneous collection (see also article III in my dissertation). Then I discuss in some more detail connections between philosophical practice and the specific approach to scientific investigation that I have above described as theoretical modeling (see also articles I-III in my dissertation). Third, I present the mechanistic approach to explanation as a template for interdisciplinary integration as well as a heuristic division of labor between different scientific disciplines studying social coordination (see also section 3c above and article V of my dissertation).

4a) In defense of methodological naturalism

Methodological naturalism in the philosophy of social science is often distinguished from ontological naturalism, or the idea that "social phenomena are natural in the sense of not transcending the contingent regularities studied by science" (Ross 2021, 122). Ross (2021, 121) appropriately describes this view as "bland because its denial seems quaint at best, if not outright unhinged, after a century and a half of development in the social sciences"). Ross (2021, 122) goes on to assert that because it "merely denies a thesis that no contemporary scientist takes seriously, it implies almost nothing in the way of a positive thesis about the ontological status of society and sociality". While some philosophers still make a point of the fact that their accounts are ontologically naturalistic (e.g. Searle 2010, 4), most serious philosophical discussions of social ontology during recent decades have focused on more specific questions, such as whether social properties are reducible to the properties (and relations of properties) of individuals—an issue that has often been discussed under the rubric of ontological individualism (Epstein 2015; Zahle&Collin 2014). In asking this question, researchers in social ontology have often presupposed that we have a relatively stable and non-controversial understanding of what individual-level (intentional) properties are—an assumption that I challenge in articles I and II of my dissertation—and that the ontological levels that are relevant to the explanation of social facts are independent of the pragmatic context of explanation (see Ylikoski&Kuorikoski 2010).

Methodological naturalism, like ontological naturalism, is sometimes described in negative rather than in positive terms—i.e. in terms of what it denies, rather than in terms of what it is committed to. Thus Ronald Giere (2008, 215) says that "the naturalist project for examining

knowledge claims in various fields of philosophical investigation rejects claims to special forms of logical and philosophical analysis, preferring to employ fundamentally the same tools employed by the relevant scientists themselves". More constructively, Giere (2008, 216-218) then goes on to mention examples of the types of tools that methodological naturalists can use, such as evolutionary models of scientific change (e.g. Hull 1988; Kuhn 1962), cognitive accounts of scientific discovery and reasoning (e.g. Magnani 1999; Thagard 2012), and sociology of science (e.g. Collins&Evans 2007). Although resolutely naturalist in his methodological commitments, Giere (2008, 214) identifies a paradox in attempting to justify methodological naturalism in terms of some a priori argument for the epistemic primacy of science over non-science, preferring to justify methodological naturalism by an inductive argument from the past successes of science. While Giere's (2008) methodological discussions are primarily motivated by concerns with the natural sciences, especially physics, one of the leading proponents of naturalism in the philosophy of social science, Harold Kincaid (2012, 391), describes its repercussions as follows:

> "On this view, philosophy of social science and social science itself are continuous, philosophy has no special knowledge or tools that only it can provide, philosophy of social science has to be intimately connected with real social research, philosophy of science can and should try to be of use to social scientists themselves, and social phenomena are susceptible to the broad methods of science in general. On the other end is the view that there are deep, eternal philosophical questions raised by the social sciences that can be and sometime can only be answered by philosophical reflection independent of and prior to the details of social research practices." (Kincaid 2012, 391)

The view that Kincaid defends highlights continuities between philosophical investigation and scientific practice, but it does not entail *methodological monism*, or the idea that there is one correct scientific method that is common to all of science (e.g. the experimental method). Rather, it allows for the possibility that there may be many different and mutually irreducible forms of scientific inquiry that are used across the broad range of sciences—as well as the claim (which I defend in articles I-III of my dissertation) that some of these forms of scientific inquiry may be closer to philosophical investigation than others. Indeed, while methodological naturalism has become an increasingly popular view to advocate during recent decades, mere endorsement of methodological naturalism does not tell us very much about the types of methodological commitments that a philosopher maintains, because of the heterogeneous nature of scientific practice (see section 3b above and article III of my dissertation). In my dissertation, I go beyond a blanket endorsement of methodological naturalism by identifying *specific* continuities between philosophical studies of shared intentionality and particular forms of scientific practice. This requires going into more detail about the particular respects in terms of which I take philosophical investigation to be comparable to scientific practice, while also accepting that some of my claims may not be applicable to other fields of philosophical investigation than the study of shared intentionality. Of course, this claim to *local* as opposed to *global applicability* is entirely in the spirit of Giere's (2008) and Kincaid's (2012) more comprehensive naturalist agendas.

The idea of methodological naturalism is an all-encompassing motivating force throughout my dissertation, and it is not reasonable to attempt to individuate all of the naturalistic claims that I make in my dissertation. Indeed, my dissertation as a whole aims at *demonstrating* the feasibility of methodological naturalism in the particular fields of philosophical investigation that I am concerned with, instead of praising it, or

presenting non-contextual and general arguments for it. However, article III of my dissertation does contain a relatively general discussion of methodological naturalism with applications to the specific field of social ontology. Moreover, article I of my dissertation is obviously penetrated by a naturalist spirit in comparing different strategies for modeling intentional agency in analytic philosophy to similar modeling strategies in science (e.g. agent-based modeling, analogical modeling, and mathematical modeling). Finally, article IV of my dissertation is a naturalistic study of minimalist approaches to shared intentionality that relies on the heuristics of decomposition and localization as an approach for finding out about mechanisms. After completing this dissertation, I have continued work on naturalistic approaches to philosophical methodology in a joint manuscript with Tuukka Kaidesoja (Sarkia&Kaidesoja 2022) concerning model-construction and inference to the best explanation as alternative naturalistic approaches to social ontology.

4b) Conceptual analysis as model-construction

The idea that methodological discussions of scientific modeling and model-construction in science can provide an instructive perspective on philosophical practice is of a relatively recent origin, and has only been explored in a few domains of philosophical investigation to this day. Notably, Peter Godfrey-Smith (2006b; 2012) and L.A. Paul (2012) have argued that certain branches of analytic metaphysics, such as the work of David Lewis on possible worlds and Humean supervenience, can be understood as forms of model-construction. Timothy Williamson (2017) has applied similar ideas to the domains of epistemology and philosophy of language, including the formal mathematical approach of Bayesian epistemology and the famous philosophical Gettier thought experiment. And Godfrey-Smith (2005) and Heidi Maibom (2003) have argued that certain important parts of ordinary folk psychological action

understanding, as well as philosophical approaches building on the conceptual analysis of our ordinary framework of intentional agency, can be understood as involving forms of theoretical modeling. However, perhaps the first contemporary philosopher to identify a significant connection between philosophical investigation and scientific model-construction, in a series of papers going as far back as the 1970s, was William Wimsatt (see also Grice 1974-5):

> "Scientists use idealizations, often conflicting ones, for various ends…. Different sets of false assumptions—idealizations—play crucial roles in teasing apart different aspects of the causal structure of our world and permeate model-building, which is the cutting edge of theory construction…. Philosophers use idealizations too—to simplify and generalize analyses, to abstract away from particular details peripheral to the point at hand, and often to urge certain norms of behavior… These idealizations too are models, though we rarely act as if they were. Calling them "constitutive ideals" or suggesting a special normative or generative role doesn't hide the fact that they embody assumptions and conditions. It is sensible to ask whether and how well these assumptions are realized in that part of the world they are applied to." (Wimsatt 2007, 16)

While Wimsatt emphasizes the roles of idealizations and abstractions (understood as deviations from the "real" nature of things—see Mäki 2020; Thomson-Jones 2005) in theoretical modeling, others have paid more attention to the indirect nature of model-based representation (e.g. Godfrey-Smith 2005), the structure of our knowledge about theoretical models (e.g. Maibom 2003), and the type of scientific progress that is involved in the development of ever better (in the sense of more accurate or inferentially more powerful) models (Williamson 2017). What is common to all of these approaches is that they are conducive to the type of pluralism about philosophical methodology that has

already been recognized by many philosophers as an essential aspect of scientific practice. In particular, they recognize that there may be many different (even mutually incompatible) models of the same phenomenona that play complementary roles in "teasing apart different aspects of the causal structure of our world" (Wimsatt 2007, 16). Moreover, they recognize that theoretical models need not represent all aspects of their targets truthfully, but may deliberately distort some aspects of the world in order to bring others into sharper relief, shifting the focus of evaluation from truth to other standards of epistemic success (e.g. simplicity, predictive power, and tractability). Although theoretical modeling and model-construction is not all of science, nor is all of philosophy engaged with theoretical modeling and model-construction, considering these forms of scientific theorizing can aid us in understanding the types of methodological practices that are inherent to particular fields of philosophical study.

My most important claims with respect to the relation between conceptual analysis in philosophy and model-construction are the following. First, in article I of my dissertation I argue that philosophers can use several different methodological strategies for constructing models of intentional agency, which is a central notion both within philosophy and in the behavioral and social sciences. In article II of my dissertation, I argue that different philosophical analyses of shared intentionality can be understood as making up a family of models of socially coordinated behavior, which build on a common understanding of basic forms of individual intentionality in our ordinary framework of agency, but which make different extensions and elaborations to this framework in order to accommodate the aspects of shared intentionality that they regard as particularly salient. And in article III of my dissertation, I argue that many philosophical accounts of social ontology, especially those dealing with social facts that are grounded by (Epstein 2015; Schaffer 2016) or supervenient on (Kim 1984; Kincaid 1986; List&Pettit 2011)

central forms of shared intentionality, can be understood as involving theoretical modeling and model-construction. Moreover, I argue that a model-based approach to social ontology can respond to naturalistic concerns about the relevance of philosophical approaches to social ontology relative to the social sciences, while also maintaining a reasonable degree of relative independence for philosophical investigation.

4c) Interdisciplinary integration through mechanistic explanation

The image of science that is conveyed in many university textbooks and in the popular media is that of almost instantaneously complete scientific models and theories coming about as a result of rare insights (e.g. "the heureka moment") or exceptional individuals (à la Albert Einstein or Isaac Newton). However, much ongoing scientific investigation is in fact concerned with the formulation and revision of relatively incomplete and defeasible mechanism sketches, rather than the types of comprehensive models and theories that university textbooks are replete with (Craver and Darden 2013, 31). This is also the case for the study of social coordination, where our understanding of its underlying cognitive (as well as ecological, environmental, social, etc.) mechanisms is still relatively poor or incomplete. For example, the cognitive scientists Thomas Dolk and his colleagues (Dolk et a. 2014, 1) testify that it "is unclear… whether processing information about other people and their activities requires special, dedicatedly "social" mechanisms… or whether universal information-processing mechanisms are sufficient". In particular, some cognitive scientists have argued that human social coordination requires irreducible 'shared task representations' (Sebanz et al. 2006; Gallotti&Frith 2013)—which have sometimes been referred to by philosophers as "we-intentions" (Searle 1990; Tuomela&Miller 19988)—where the actions of one's collaborative partner are

represented under a common executive format with one's own goal-directed activities. On the other hand, many of the cognitive mechanisms that we use for coordinating with others seem to be shared with other species in the animal kingdom. This is so not only for mirror neurons, which were first discovered in the frontal premotor areas of macaque monkeys (Rizzolatti&Sinigaglia 2010), but also for many other mechanisms of emergent and planned coordination (see Knoblich et al. 2011; Vesper et al. 2016).

The mechanistic framework of explanation in the philosophy of science provides a convenient way of framing questions about how complete or comprehensive our current understanding of the nature and mechanisms of social coordination is. For example, we may ask if we know what forms of social coordination there are, what types of cognitive mechanisms dispose agents to engage in such forms of social coordination, how robust our understanding of those mechanisms is, and how their operations are modulated by the natural and social environments that such agents populate. On the other hand, lack of evidence about mechanisms for particular forms of social coordination may prompt questions about whether those forms of social coordination constitute relevant categories for scientific investigation, whether their properties should be characterized differently, or whether their boundaries are less robust than one might have at first have imagined. To take up just one example that is discussed in article V of my dissertation, several proponents of the so-called *Vygotskian intelligence hypothesis* in evolutionary anthropology have suggested that many highly abstract and complex forms of human culture, such as "money, marriage, and government, which only exist due to the shared practices and beliefs of a group" (Tomasello et al. 2005, 680) are dependent on special forms of social intelligence that are unique to the human species. However, I argue in my paper that failure to identify species-distinctive cognitive mechanisms for such forms of social intelligence undermines the plausibility

of the Vygotskian intelligence hypothesis as a general account of human cultural uniqueness. My investigation indicates that interdisciplinary integration need not always be a smooth or gradual process, but may require substantial revision of the assumptions and hypotheses that have been put forth in some of the research programs to be integrated.

5. Brief introduction to the articles in my dissertation

5a) Modeling the social world through individual and shared intentionality (articles I-III)

The articles in this part of my dissertation form a tightly knit package that is concerned largely with questions of philosophical methodology, proceeding from the study of individual intentionality (Sarkia 2021a) through shared intentionality (Sarkia 2022a) to the ontology of the social world (Sarkia 2021b). In each of these domains, I defend a model-based, pluralistic, and pragmatic approach to the domains of philosophical study in question—an approach that is based on relevant analogies to similar practices of scientific modeling and model-construction.

The first article in part A of my dissertation, *Modeling Intentional Agency: a neo-Gricean Account*, draws a distinction between the substantive features of philosophical models of intentional agency—such as the widely used belief-desire-intention (BDI)-framework (Bratman 1987)—and the methodological strategies by means of which such models are constructed. Moreover, it analyzes three different strategies for modeling intentional agency in analytic philosophy—which I call Gricean modeling, analogical modeling, and theoretical modeling—and compares them to similar modeling strategies in science, such as agent-based modeling, use of model organisms, and mathematical modeling. To my

knowledge, the idea that *similar* models of intentional agency could be constructed by way of many *different* modeling strategies, prompting questions of robustness and complementarity to arise, has not before been defended in the vast literature on the philosophy of action, where philosophers have generally assumed a type of unity of method ("conceptual analysis") that, in light of my investigation, does not seem justified.

The second article in part A my dissertation, *A Family of Models of Shared Intentionality*, argues that many philosophical debates about whether shared intentionality is reducible to individual intentionality or not come down to contrasting background views about individual intentionality, rather than to contrasting views about the nature of shared intentionality *per se*. Moreover, it argues that different philosophical accounts of shared intentionality can be understood as a family of models, which draw on a common background of core assumptions in our ordinary framework of intentional agency (e.g. that all mental states have a direction of fit), but make contrasting extensions and elaborations to this core in order to accommodate the aspects of shared intentional action that they regard as particularly salient. My view is supported by arguments to the effect that our ordinary framework of agency is itself structured as a family of models (Godfrey-Smith 2005), and parallels to different models of rationality in microeconomics (Hausman 1992; Ross 2010).

The third article in part A my dissertation, *A Model-Based Approach to Social Ontology*, presents the *naturalist's conundrum* as a methodological challenge for naturalistically oriented philosophers of social ontology, which consists of the twin desiderata of relevance for day-to-day social scientific inquiry and safeguarding a degree of relative independence for philosophical research (primarily in relation to empirical social science). My response to the naturalist's conundrum (which I do not claim to be

the only possible response) draws centrally on the two-stage picture of scientific theorizing that is characteristic of theoretical modeling as a distinctive approach to scientific theorizing, which proceeds by first constructing a hypothetical model (which may be in many respects abstract and idealized relative to the real-world systems that the model is supposed to represent), and then using the model to draw inferences about its real-world targets (Godfrey-Smith 2006; Weisberg 2007). Although there may in practice be significant back-and-forth motion between the two stages of the model-based enterprise, as a model is adjusted and calibrated in order to better capture salient features of its targets, this solution guarantees a reasonable degree of independence to the activity of model-construction (which also philosophers—but not only philosophers—may partake in). My solution to the naturalist's conundrum is illustrated in this paper through a detailed case study of Raimo Tuomela's philosophical account of the we-perspective.

5b) Mechanisms of social coordination from an interdisciplinary perspective (articles IV-VI)

This part of my dissertation takes a broader interdisciplinary approach, focusing on the connections between different disciplinary research programs studying social coordination. The mechanistic approach to explanation plays an important role in several articles in this part of my dissertation, grounding a schematic division of labor between behavioral and cognitive scientists studying social coordination, and providing a blueprint for how their theoretical outputs can be integrated with one another (when they are mutually compatible—which may not always be the case). While the first article in part B of my dissertation is concerned with minimalist accounts of shared intentional action in cognitive science and in philosophy, the latter two articles address

interdisciplinary relations between the behavioral, cognitive, and social sciences.

The first article in part B of my dissertation, *Minimalisn and Maximalism in the Study of Shared Intentional Action*, is concerned with minimalist accounts of shared intentional action in cognitive science and in philosophy. Several minimalist approaches to shared intentional action have been motivated by work in developmental psychology on the ontogeny of human capacities for meta-representation, which do not seem to develop fully (although they are present in more rudimentary form) until around four years of age. However, given that infants engage in some joint activities even before this age, many minimalists have argued that there must be some forms of shared intentional action that do not require the types of meta-representational capacities that are demanded by many standard accounts of shared intentional action (e.g. Bratman 1999; 2014). Other minimalists have presented their accounts as analyses of the cognitive mechanisms that make the execution of more complex shared intentional actions possible. In my paper, I pull apart these "complementarist" and "constitutionalist" versions of the minimalist challenge, argue that they have not been sufficiently distinguished in the literature, and connect them to the heuristics of decomposition and localization as an approach to mechanism discovery (Bechtel&Richardson 2010).

The second article in part B of my dissertation, *Mechanistic Explanation, Interdisciplinary Integration, and Interpersonal Social Coordination*, makes use of the distinction between phenomenal decomposition and mechanistic decomposition (Glennan 2017; Craver&Darden 2013) to carve out a heuristic division of labor between behavioral and cognitive scientists studying interpersonal social coordination, and to indicate how their theoretical insights can be integrated with one another. Thus the mechanistic approach to explanation functions both as a framework for

mechanism discovery, and as a framework for *interdisciplinary integration*, given the hypotheses, models, and theories that the process of mechanism discovery results in. In my article, I consider potential tensions in the process of interdisciplinary integration, in addition to situations, where integration proceeds smoothly. To this end, I consider mechanistic evidence (which I argue to be inconclusive) in favor of the Vygotskian intelligence hypothesis, according to which human beings possess species-distinctive forms of cultural intelligence that other primates do not possess. My article argues that a modified version of the Vygotskian intelligence hypothesis that I refer to as the *social motivation hypothesis*, following Godman (2014; Godman et al. 2014) is better supported by the empirical data, but has narrower explanatory scope, forcing proponents of the Vygotskian intelligence hypothesis to give up some of their more ambitious theoretical aims.

The third and final article in part B of my dissertation, *Mechanistic Explanations in the Cognitive Social Sciences: Lessons from Three Case Studies*, discusses interdisciplinary relations between the cognitive sciences and the social sciences. In this joint paper with Tuukka Kaidesoja and Mikko Hyyryläinen, we consider three social scientific research programs, which aspire to integrate insights from the cognitive sciences. Our case studies deal with the phenomena of social coordination, transactive memory systems, and ethnicity. The mechanistic framework of explanation plays an important role in helping us understand what types of insights can be gleaned from these attempts at crossing disciplinary boundaries, and where more work remains to be done. To answer these questions, we use the intuitive visual metaphors of *looking at* the phenomenon to be explained, *looking down* at the entities and activities that give rise to the phenomenon, *looking around* at the ways in which these entities and activities are organized, and *looking up* at how the phenomenon is situated in the broader environment that it is embedded in. We also use mechanistic philosophy of science to question pre-reflective

distinctions about cognitive and social domains or 'levels of reality', and to illustrate how social scientists can contribute to research programs in cognitive science, in addition to benefiting from cognitive scientific inputs.

References

Acemoglu, D., & Robinson, J. (2012). *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. London: Profile.

Amadae, S. (2015). *Prisoners of Reason: Game Theory and the Neoliberal Political Economy*. Cambridge: Cambridge University Press.

Aoki, M. (2001) *Toward a Comparative Institutional Analysis*. Cambridge, MA: MIT Press.

Apostel, L. (Ed.). (1972). *Interdisciplinarity: Problems of Teaching and Research in Universities*. Paris: Organization for Economic Cooperation and Development.

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

Aumann, R. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8, 6-19.

Aydinonat, N. E., & Ylikoski, P. (2018). Three conceptions of a theory of institutions. *Philosophy of the Social Sciences*, *48*(6), 550-568.

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, 55(4), 1278–1289.

Bach, K. (1995). Terms of Agreement. *Ethics*, *105*(3), 604–612.

Bates, R., Greif, A., Levi, M., Rosenthal, J., & Weingast, B. (2020). *Analytic Narratives*. Princeton, NJ: Princeton University Press.

Bateson, G. (1973). *Steps to an Ecology of Mind*. Frogmore, U.K.: Paladin.

Bechtel, W. & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of the Biological and Biomedical Sciences*, 36: 421–441.
Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press.
Bhaskar, R. (1979/1998). *The Possibility of Naturalism: a Philosophical Critique of the Contemporary Human Sciences. 3rd ed.* London: Routledge.
Bicchieri, C. (2006). *The Grammar of Society*. Cambridge: Cambridge University Press.
Boyd, R. (1991) Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds. *Philosophical Studies*, 61, 127–148.
Boyd, R. & Richerson, P. (2005) *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.
Bowles, S., and Gintis, H. (2011). *A Cooperative Species. Human Reciprocity and its Evolution*. Princeton: Princeton University Press.
Braddon-Mitchell, D., & Jackson, F. (2007). *Philosophy of mind and cognition. An introduct°n* (2nd ed.). Oxford: Blackwell.
Bratman, M. (1987). *Intention, Plans and Practical Reason*. Cambridge, MA: Harvard University Press.
Bratman, M. (1999). *Faces of Intention*. Cambridge: Cambridge University Press.
Bratman, M. (2014) *Shared Agency: a Planning Theory of Acting Together*. Oxford: Oxford University Press.
Brentano, F. (1874). *Psychology from an Empirical Standpoint*, London: Routledge and Kegan Paul.
Broadbent, A. (2013). *Philosophy of Epidemiology*. Palgrave Macmillan.
Butterfill, S., & N. Sebanz. (2011). Joint action: what is shared? *Review of Philosophy and Psychology*, 2(2), 137–146.
Butterfill, S. (2012). Joint action and development. *The Philosophical Quarterly*, 62(246), 23-47.

Butterfill, S. (2020). *The Developing Mind: a Philosophical Introduction*. London: Routledge.
Bzdok, D., & Dunbar, R. (2020). The neurobiology of social distance. *Trends in Cognitive Sciences*, *24*(9), 717-733
Carpenter, M. & Svetlova, M. (2016) Social development. In: Hopkins, B., Geangu, E. & Linkenauer, S. (Eds.) *Cambridge Encyclopedia of Child Development* (pp. 415-423). Cambridge: Cambridge University Press.
Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.
Cartwright, N. (2007) *Hunting Causes and Using Them: Approaches in Philosophy of Economics*, Cambridge: Cambridge University Press.
Caston, Victor (1998). Aristotle and the problem of intentionality. *Philosophy and Phenomenological Research*, 58(2), 249-298.
Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality & Social Psychology*, 76(6), 893–910.
Christensen, W., & Michael, J. (2016). From two systems to a multi-systems architecture for mindreading. *New Ideas in Psychology*, *40*, 48-64.
Collins, H. & Evans, R. (2007). *Rethinking Expertise*. Chicago, IL: University of Chicago Press.
Colombo, C. & Guala, F. (2021) Coordination without meta-representation. *Philosophical Psychology* . Forthcoming.
Crane, T. (1998). Intentionality as the mark of the mental. In A. O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind*, Cambridge: Cambridge University Press.
Craver, C. (2002). Structures of scientific theories. In P. Machamer & M. Silberstein (Eds.), *Blackwell guide to the philosophy of science* (pp. 55-79). Oxford: Blackwell.
Craver, Carl F. (2003). The making of a memory mechanism. *Journal of the History of Biology*, 36(1), 153-95.
Craver C. & Darden L. (2013). *In Search of Mechanisms: Discoveries Across the Life Sciences*. Chicago, IL: University of Chicago Press.

Da Costa, N. C., & French, S. (1990). The model-theoretic approach in the philosophy of science. *Philosophy of Science*, *57*(2), 248-265.
Da Costa, N., & French, S. (2003). *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford: Oxford University Press.
Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685-700.
Davidson, D. (1980/2001). *Essays on Actions and Events. 2nd. Edition.* Oxford: Clarendon Press.
Davis, J., Redshaw, J., Suddendorf, T., Nielsen, M., Kennedy-Costantini, S., Oostenbroek, J., & Slaughter, V. (2021). Does Neonatal Imitation Exist? Insights From a Meta-Analysis of 336 Effect Sizes. *Perspectives on Psychological Science*, *16*(6), 1373–1397.
Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain sciences*, *1*(4), 568-570.
Dolk, T., Hommel, B., Colzato, L., Schütz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. *Frontiers in Psychology*, 5, 1-10.
Downes, S. (2011). Scientific models. *Philosophy Compass*, 6(11), 757-764.
Dumas, G., & M. Fairhurst. (2021). Reciprocity and alignment: quantifying coupling in dynamic interactions. *Royal Society Open Science*, 8(5), 210138.
Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6), 469–493.
Dunbar, R. (2009). The social brain hypothesis and its implications for social evolution. *Annals of Human Biology*, 36(5), 562–572.
Dupré, J. (1993) *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.
Epstein, B. (2015). *The Ant Trap. Rebuilding the Foundations of the Social Sciences*. New York, NY: Oxford University Press.
Epstein, B. (2018). Social ontology. *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL =

<https://plato.stanford.edu/archives/sum2018/entries/social-ontology/>.
Eronen, M. (2015). Robustness and reality. *Synthese,* 192, 3961–3977.
Frith C., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, 63, 287-313.
Frodeman, R. (2017). The future of interdisciplinarity. In Frodeman et al. (2017), 1-8.
Frodeman, R., Klein, J., & Pacheco, R. (2017). *The Oxford Handbook of Interdisciplinarity. 2nd. Ed.* Oxford: Oxford University Press.
Gallotti, M. & Frith, C. (2013). Social cognition in the we-mode. *Trends in the Cognitive Sciences*, 17(4), 160-165.
Gallotti, M., Fairhurst, M., & Frith, C. (2017). Alignment in social interactions. *Consciousness and Cognition*, 48, 253-261.
Giere, R. (1988). *Explaining Science: a Cognitive Approach.* Chicago, IL: University of Chicago Press.
Giere, R. (1999). *Science Without Laws.* Chicago, IL: University of Chicago Press.
Giere, R. (2004). How models are used to represent reality. *Philosophy of Science*, 71(5), 742-752.
Giere, R. (2008). Naturalism. In S. Psillos & M. Curd (Eds.). *The Routledge Companion to the Philosophy of Science* (pp. 308-310). London: Routledge.
Gilbert, M. (1990). Walking together: a paradigmatic social phenomenon. *Midwest Studies in Philosophy*, 15(1), 1-14.
Gilbert, M. (2013). *Joint Commitment. How We Make the Social World.* Oxford: Oxford University Press.
Gintis, H. (2007). The Evolution of Private Property. *Journal of Economic Behavior & Organization*, 64(1), 1-16.
Gintis, H. (2009). *The Bounds of Reason.* Princeton: Princeton University Press.
Glennan S. (2017). *The New Mechanical Philosophy.* Oxford: Oxford University Press.

Glennan, S. & Illari, P. (2018) *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.
Godfrey-Smith, P. (2003). *Theory and Reality*. Chicago: Chicago University Press.
Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosopher's Imprint*, 5(6), 1-16.
Godfrey, Smith. P. (2006a). The strategy of model-based science. *Biology and Philosophy*, 21(5), 725-740.
Godfrey-Smith, P. (2006b). Theories and models in metaphysics. *The Harvard Review of Philosophy*, 14(1), 4-19.
Godfrey-Smith, P. (2012). Metaphysics and the philosophical imagination. *Philosophical Studies*, 160, 97-113.
Godman, M. (2020). *The Epistemology and Morality of Human Kinds*. London: Routledge.
Gold, N., & Sugden, R. (2007). Collective intentions and team agency. *Journal of Philosophy*, 104(3), 109–137.
Goldman, A. (2008) *Simulating Minds: the Philosophy, Psychology and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Gomez-Lavin, J., & Rachar, M. (2019). Normativity in joint action. *Mind & Language*, 34(1), 97-120
Gomez-Lavin, J. & Rachar, M. (2021). Why we need a new normativism about collective action. *Philosophical Quarterly*, 72(2), 478-507.
Goodman, N. (1976). *Languages of Art*. Indianapolis: Hackett.
Gopnik, A. & Meltzoff, A. (1997). *Words, Thoughts, Theories*. Cambridge, MA: MIT Press.
Greif, A. (2006). *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge: Cambridge University Press.
Greif, A. & Kingston, C. (2011). Institutions: Rules or Equilibria? In A. Greif & C. Kingston (Eds.), *Political Economy of Institutions, Democracy and Voting* (pp. 13-43). Dordrecht: Springer.
Greenwood, J. (2015). *A Conceptual History of Psychology. 2nd Ed.* Cambridge University Press.

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology*, 45(5), 1430-43.
Guala, F. (2007). The philosophy of social science: metaphysical and empirical. *Philosophy Compass*, 2(6), 954-80.
Guala, F. (2016). Naturalism and anti-naturalism in the philosophy of social science. In P. Humphreys (Ed.), *The Oxford handbook of philosophy of science* (pp. 43-61). Oxford: Oxford University Press.
Hacking, I. (1983) *Representing and Intervening.* Cambridge: Cambridge University Press.
Hakli, R. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7(2-3), 286-297.
Hakli, R., Miller, K., & Tuomela, R. (2011). Two kinds of we-reasoning. *Economics and Philosophy*, 26, 291-320.
Hamann, K., Warneken, F., Greenberg, J., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, 476(7360), 328–31.
Hardin, G. (1968). The tragedy of the commons. *Science*, 162, 1243-1248.
Hare, B. & Tomasello, M. (2004). Chimpanzees are more skillful in competitive than in cooperative cognitive tasks. *Animal Behaviour, 68, 571-81.*
Hammerstein, P. & R. Noë. 2016. Biological Trade and Markets. *Philosophical Transactions of the Royal Society B* 371 (1687).
Hare, B. (2017). Survival of the friendliest: Homo sapiens evolved via selection for prosociality. *Annual Review of Psychology*, *68*, 155-186.
Harman, G. (1986). *Change in View. Principles of Reasoning.* Cambridge (MA): MIT Press.
Harris, C.B., A.J. Barnier, J. Sutton, and P.G. Keil, 2014, Couples as Socially Distributed Cognitive Systems: Remembering in Everyday Social and Material Contexts. *Memory Studies*, 7(3), 285–297.

Harsanyi, J. (1975). The tracing procedure. *International Journal of Game Theory*, *4*, 61–94.
Hausman, D. (1992). *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.
Hedstrom, P. (2005). *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press
Heinonen, M. (2016). Minimalism and maximalism in the study of shared intentional action. *Philosophy of the Social Sciences*, 46(2), 168-188.
Hempel, C. (1966) *Philosophy of Natural Science*. Englewood Cliffs, CA: Prentice-Hall.
Henrich, J. & Muthukrishna, M. (2021) The origins and and psychology of human cooperation. *Annual Reviews in Psychology*. Forthcoming.
Hindriks, F. (2009) Constitutive Rules, Language, and Ontology. *Erkenntnis,* 71, 253-75.
Hindriks, F. & Guala, F. (2019). The functions of institutions: etiology and teleology. *Synthese*, 198(3), 2027-2043.
Hrdy, S. (2009) *Mothers and Others. The Evolutionary Origins of Mutual Understanding*. Cambridge, MA: Harvard University Press.
Hull, D. L. (1988). *Science as a Process: an Evolutionary Account of the Social and Conceptual Development of Science*. Chicago, IL: University of Chicago Press.
Jacobs, J. A. (2017). The need for disciplines in the modern research university. In Frodeman et al. (2017) , 35-40.
Jankovic, M., & Ludwig, K. (2017) Collective intentionality. In L. McIntyre & A. Rosenberg (Eds.), *Routledge Companion to Philosophy of Social Science*. London: Routledge:
Jarvie, I. C., & Zamora-Bonilla, J. (2011). *The Sage Handbook of the Philosophy of Social Sciences*. Thousand Oaks, CA: Sage.
Kachel, U., & Tomasello, M. (2019). 3-and 5-year-old children's adherence to explicit and implicit joint commitments. *Developmental Psychology*, 55(1), 80-88.

Kaidesoja, T. (2013). *Naturalizing Critical Realist Social Ontology*. London: Routledge.
Kim, J. (1984). Concepts of Supervenience. *Philosophy and Phenomenological Research*, 45(2), 155-17
Kincaid, H. (1986). Reduction, Explanation, and Individualism. *Philosophy of Science*, 53(4), 492-513.
Kincaid, H. (2012). How should philosophy of social science proceed? *Metascience*, 21(2), 391-394.
Klein, J. T. (2017). Typologies of interdisciplinarity. In Frodeman et al. (2017), 21-34.
Knoblich, G., Butterfill S and Sebanz N. (2011). Psychological research on joint action: theory and data. In B. Ross (Ed.) *Psychology of Learning and Motivation vol. 51* (pp. 59-101). Burlington: Academic Press.
Koskinen, I., & Rolin, K. (2019). Scientific/intellectual movements remedying epistemic injustice: The case of Indigenous studies. *Philosophy of Science*, 86(5), 1052–1063.
Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for the Philosophy of Science*, 61, 541-567.
Kuorikoski, J., & Pöyhönen, S. (2012). Looping Kinds and Social Mechanisms. *Sociological Theory*, *30*(3), 187–205.
Lakin, J., & Chartrand, T. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14, 334-339.
Leyton-Brown, K. & Y. Shoham. (2008). *Essentials of Game Theory. A Concise, Multidisciplinary Introduction*. Morgan&Claypool.
Levins, R. (1966). The strategy of model building in population biology. *American scientist*, *54*(4), 421-431
Lewis, D.(1969). *Convention: A Philosophical Study*. Cambridge, MA: Wiley-Blackwell.

List, C., and Pettit, P. (2011). *Group Agency. The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
Longino, H. 2013. *Studying Human Behavior. How Scientists Investigate Aggression and Sexuality*. Chicago, IL: University of Chicago Press.
List, C. & Pettit, P. (2011). *Group Agency. The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford University Press.
Ludwig, K. (2016). *From Individual to Plural Agency*. Oxford: Oxford University Press.
Ludwig, K. (2017). *From Plural to Institutional Agency*. Oxford: Oxford University Press.
Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751-19756.
Machamer, P., Darden, L. & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1-25.
Machery, E. (2017). *Philosophy Within its Proper Bounds*. Oxford: Oxford University Press.
Macleod, M. (2018). What makes interdisciplinarity difficult? *Synthese*, 195, 697-720.
Magnani, L., Nersessian, N., & Thagard, P. (Eds.). (1999). *Model-Based Reasoning in Scientific Discovery*. New York: Springer.
Maibom, H. (2003). The Mindreader and the Scientist. *Mind&Language*, 18(3), 296-315.
Maynard-Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
Meltzoff, A. & Moore, K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198(4312), 75–78.
Menger, C. (1892). On the Origin of Money. *Economic Journal*, 2, 239–55.
Michael, J. (2011). Shared emotions and joint action. *Review of Philosophy and Psychology*, 2, 355-373.

Michael, J. & Pacherie, E. (2015). On commitments and other uncertainty reduction tools in joint action. *Journal of Social Ontology*, 1(1), 89-120.
Michael, J., Sebanz, N., & Knoblich, G. (2016). The sense of commitment: a minimal approach. *Frontiers in psychology*, *6*, 1968.
Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63, 81-97.
Miller, S. (2001). *Social Action: a Teleological Account*. Cambridge: Cambridge University Press.
Miller, S. (2010). *The Moral Foundations of Social Institutions: A Philosophical Study*. Cambridge: Cambridge University Press.
Morgan, M. (2012). *The World in the Model: How Economists Work and Think*. Cambridge: Cambridge University Press.
Mäki, U. (2009). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis*, 70 (1), 29–43.
Mäki, U. (2016). Philosophy of interdisciplinarity: what? Why? How? *European Journal for Philosophy of Science*, 6, 327-342.
Mäki, U. (2020). Puzzled by idealizations and understanding their functions. *Philosophy of the Social Sciences*, 50(3), 215-237.
Mäki, U. (2020b). Reflections on the Ontology of Money. *Journal of Social Ontology*, *6*(2), 245-263.
Nagatsu, M., Davis, T., DesRoches, C. T., Koskinen, I., MacLeod, M., Stojanovic, M., and Thorén, H. (2020). Philosophy of Science for Sustainability Science. *Sustainability Science,* 15, 1807-1817.
North, D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
Northcott, R. & Alexandrova, A. (2015). The prisoner's dilemma doesn't explain much. In Peterson (2015).
O'Brien, L. (2015). *Philosophy of Action*. Hampshire: Palgrave Macmillan.
Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge Uni-versity Press.

Pacherie, E. (2013). Intentional joint agency: shared intention lite. *Synthese*, 190, 1817-1839.
Paul, L. (2012). Metaphysics as modeling: the handmaiden's tale. *Philosophical Studies*, 160(1), 1-29.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences*, *1*(4), 515-526.
Quinton, A. (1975-76). Social objects. *Proceedings of the Aristotelian Society*, 76, 1-27.
Rakoczy, H. (2017) The development of individual and shared intentionality. In: Kiverstein J (Ed.), *The Routledge Handbook of the Philosophy of the Social Mind* (pp. 139-151). London: Routledge..
Richardson, M., Marsh, K., Isenhower, R., Goodman, J. & Schmidt, R. (2007) Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, 26(6), 867-891.
Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11, 264–274.
Ross, D. (2010). The economic agent: not human, but important. In U. Mäki (Ed.), *Elsevier handbook of philosophy of science, vol. 13: economics*. Amsterdam: Elsevier.
Ross, D. (2011). Naturalism: the place of society in nature. In Jarvie and Zamora Bonilla (2011), 121-136.
Ross, D. (2012). Coordination and the foundations of social intelligence. In *The Oxford Handbook of Philosophy of Social Science*.
Roth, A. (2016). *Who Gets What and Why: the Economics of Matchmaking and Market Design*. Earmon Molan/Mariner.
Salmela, M. (2013). Shared emotions. *Philosophical Explorations*, 15(1), 33-46.
Saint-Germier, P., Paternotte, C., & Canonne, C. (2021) Joint improvisation, minimalism and pluralism about joint action. *Journal of Social Ontology*, 7(1), 97-118.

Sarkia, M., Kaidesoja, T., & Hyyryläinen, M. (2020). Mechanistic explanations in the cognitive social sciences: lessons from three case studies. *Social Science Information*, 59(4), 580-603.
Sarkia, M. (2021a). Modeling intentional agency: A neo-Gricean framework. *Synthese, 199, 7003-7030.*.
Sarkia, M. (2021b). A model-based approach to social ontology. *Philosophy of the Social Sciences. OnlineFirst.*
Sarkia, M. (2022a). A family of models of shared intention. *Under review.*
Sarkia, M. (2022b). Mechanistic explanation, interdisciplinary integration, and interpersonal social coordination. *Under review.*
Sarkia, M. & Kaidesoja, T. (2022). Two approaches to naturalistic social ontology. *Article manuscript.*
Schaffer, J. (2016). Grounding in the Image of Causation. *Philosophical Studies* 173 (1): 49-100.
Schelling, T. (1960) *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.
Schelling, T. (1978). *Micromotives and Macrobehavior.* New York, NY: W.V. Norton.
Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind.* Cambridge: Cambridge University Press.
Searle, J. (1990/2002). Collective intentions and actions. In J. Searle. 2002. *Consciousness and Language* (pp. 90-105). Cambridge: Cambridge University Press.
Searle, J. (2009). Language and social ontology. In C. Mantzavinos (ed.). *Philosophy of the Social Sciences. Philosophical Theory and Scientific Practice* (pp. 9-27). Cambridge: Cambridge University Press.
Searle, J. (2010). *Making the Social World: the Structure of Human Civilization.* New York: Oxford University Press.
Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2), 70–76.
Seeley, T. (1995). *The Wisdom of the Hive. The Social Physiology of Honey Bee Colonies.* Cambridge, MA: Harvard University Press.

Sellars, W. 1963. *Science, Perception and Reality*. Atascadero, CA: Bridgeview.
Simon, H. (1969): *The Sciences of the Artificial*. Cambridge, MA: MIT University Press.
Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, 189, 260-274.
Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.
Smit, J., Buekens, F., & Du Plessis, S. (2011). What is money? An alternative to Searle's institutional facts. *Economics and Philosophy, 27*(1), 1-22.
Spelke, E. S., Bernier, E. P., & Skerry, A. (2013). Core social cognition. In Banaji, M. & Gelman, S. (Eds.). *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us* (pp. 11-16). Oxford: Oxford University Press.
Suarez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science*, 71(5), 767-779.
Suppe, P. (1960). A comparison of the meaning and use of models in mathematics and the empirical sciences. *Synthese*, 12(2-3), 287-301.
Suppes, P. (1966). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.). *Studies in Logic and the Foundations of Mathematics* (Vol. 44, pp. 252-261). Elsevier.
Suppes, F. (1989) *The Semantic Conception of Theories and Scientific Realism*. Chicago, IL: University of Chicago Press.
Thagard, P. (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. Harvard, MA: MIT Press.
Thomson-Jones, M. (2005). Idealization and abstraction: A framework. In M. Thomson-Jones and N. Cartwright (Eds), *Idealization XII: Correcting the Model* (pp. 173-217). Amsterdam: Rodopi.
Tollefsen, D. (2005). Let's pretend! Children and joint action. *Philosophy of the Social Sciences*, 35(1), 75-97.
Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: 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, 675-691.

Tomasello, M., & Carpenter, M., (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child development*, i-152

Tomasello M. (2019) *Becoming Human: a Theory of Ontogeny*. Cambridge, MA: Harvard University Press.

Trevarthen, C. & Aitken, K. 2001. Infant intersubjectivity: research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry*, 42(1), 3–48.

Tronick, E., Als, H., Adamson, L., Wise, S., & Brazelton, B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry*, 17, 1–13.

Tuomela, R. (2007). *The Philosophy of Sociality*. New York: Oxford University Press.

Tuomela, R. (2013). *Social Ontology: Collective Intentionality and Group Agents*. New York: Oxford University Press.

Tuomela, R. & Miller, K.. (1988). We-Intentions. *Philosophical Studies*, 53(3), 367-389.

Vanderschraaf, P. (1995). Convention as Correlated Equilibrium. *Erkenntnis*, 42, 65-87.

Van Fraassen, B. (1987). The semantic approach to scientific theories. In N. Nersessian (Ed.). *The Process of Science: Contemporary Philosophical Approaches to Understanding Scientific Practice* (pp. 105-124). Dordrecht: Kluwer.

Velleman, D. (1997). How to share an intention. *Philosophy and Phenomenological Research*, 57(1), 29-50.

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., Hristova, D., Karlinsky, A., McEllin, L., Nijssen, S., Schmitz, L., & Wahn, B. (2016) Joint action: mental representations, shared

information and general mechanisms for coordinating with others. *Frontiers in Psychology*, 7, 1-7.
Warneken, F., Chen, F., & Tomasello, M. (2006) Cooperative activities in young children and chimpanzees. *Child Development*, 77(3),640-63.
Wegner, D. M. (1986). Transactive memory: a contemporary analysis of the group mind. In B. Mullen & G. Goethals (Eds.), *Theories of group behavior* (pp. 185-208). Dordrecht: Springer.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730-742.
Weisberg, M. (2007a). Who is a Modeler?. *The British Journal for the Philosophy of Science*, 58(2), 207-233.
Weisberg, M. (2007b). Three kinds of idealization. *The Journal of Philosophy*, 104(12), 639-659.
Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.
Williamson, T. 2017. Model-Building in Philosophy. In R. Blackford & D. Broderick (Eds.). *Philosophy's Future: the Problem of Philosophical Progress*, (pp.159-172). Hoboken, NJ: Wiley.
Wilson, E. (2012). *The Social Conquest of Earth*. New York: Norton.
Wimsatt, W. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13,* 219–240.
Wrangham, R. (2019). Hypotheses for the evolution of reduced reactive aggression in the context of human self-domestication. *Frontiers in Psychology*, 10, 1-11.
Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, 148, 201-219.
Zahle, J., & Collin, F. (2014). *Rethinking the Individualism-Holism Debate*. Dordrecht: Springer.