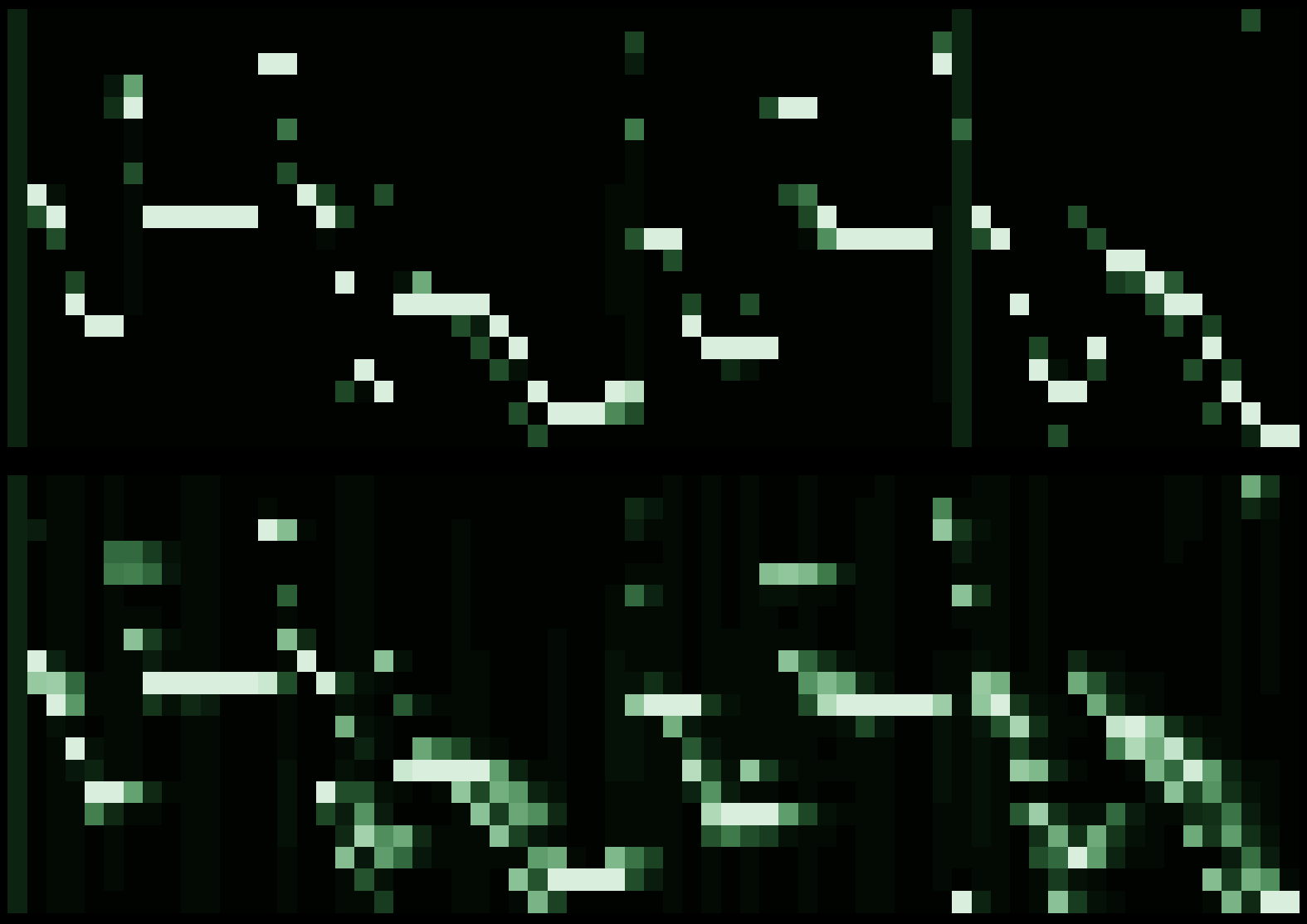
SOCIAL MOTORICS

a predictive processing model for efficient embodied communication

SEBASTIAN KAHL



SOCIAL MOTORICS

a predictive processing model for efficient embodied communication

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doktor der Ingenieurwissenschaften (Dr.-Ing.)
at the Faculty of Technology at Bielefeld University

by

SEBASTIAN KAHL

SOCIAL MOTORICS — A PREDICTIVE PROCESSING MODEL FOR EFFICIENT EMBODIED COMMUNICATION

DOI: https://doi.org/10.4119/unibi/2945718

SEBASTIAN KAHL
Social Cognitive Systems Group, Faculty of Technology,
Bielefeld University, PO Box 10 01 31, 33501 Bielefeld, Germany
skahl@uni-bielefeld.de
https://orcid.org/0000-0002-8468-2808

Von der Technischen Fakultät der Universität Bielefeld zur Erlangung des Grades eines Doktor der Ingenieurwissenschaften (Dr.-Ing.) genehmigte Dissertation.

DISSERTATION COMMITTEE:
- Prof. Dr.-Ing. Stefan Kopp (Supervisor, Bielefeld University)
- Prof. Dr. Martin V. Butz (University of Tübingen)
- Dr. Malte Schilling (Bielefeld University)
- Prof. Dr. Helge Ritter (Bielefeld University)

DATE OF SUBMISSION: 2020-03-02
DATE OF DEFENSE: 2020-08-18

This document was typeset using the typographical look-and-feel classicthesis v4.6 developed by André Miede and Ivo Pletikosić.
The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX: https://bitbucket.org/amiede/classicthesis/

The paper used in this publication meets the requirements for permanence of paper as specified in ISO 9706:1994.

We seem to be a most curious breed of dysfunctional clairvoyants,
looking forward to the theoretical possibility,
while only becoming aware of what we have actually embarked upon
once we have already started.

— Spence (2009, pp. 125-126)

The critical act in formulating computational theories turns out
to be the discovery of valid constraints on the way the world
is structured — constraints that provide sufficient information
to allow the processing to succeed.

— Marr (1980)

ACKNOWLEDGMENTS

These kinds of writings are never and should never be created alone. So, wholeheartedly, I want to thank some special people for their continuing support during this time.

First of all, I want to thank my supervisor Stefan Kopp, who always supported my crazy musings about where this work should be heading. During our discussions I always felt that he is as exited about its potential as I am. His valuable support even encouraged me to send him Whatsapp messages during off-hours when I had one or the other *heureka* moments, when something finally worked.

During my time at the Social Cognitive Systems Group I always felt at home! I want to thank my former and fellow colleagues for all the support at work, the fun nights out, the crazy chats, the many many (many) coffees and, last but not least, the breakfasts. Namely, I especially want to thank Jan Pöppel for his open ear and the fun discussions about all the technical ideas and problems I had with *the model*. Also, I want to thank Laura Hoffmann, Sonja Stange, Philipp Kulms, Hendrik Buschmeier, and Farina Freigang. Thank you for your open ears to my troubles over the years. Thanks to you this work did not break me. Over the years many colleagues came and went, but the heart of the group never changed. Dagmar Philipp represents this heart. I especially want to thank her for her invaluable support over the years. Whenever I had any question she would drop everything and try to help me.

Further special thanks go to Martin Butz who agreed to be the second reviewer of this thesis.

I also want to thank all my friends for their support, you know who you all are! You never dismissed my rants about *the model*, how it isn't working, what it should be able to do, what it finally did.

I am indebted to my family for their support. I still remember their faces when I first told them that I wanted to study Cognitive Science. Today, you know how to describe to your friends and colleagues what I do, AI and pretty pictures from neuroscience are everywhere. Thank you for believing in me and supporting me, Birgit Tennigkeit-Kahl, Gerhard Kahl, and Stephan Kahl.

Finally, I want to thank my special one. You have been there for me when nothing worked, you were excited with me when it finally did, you supported me and believed in me when I was down.
My fiancée Kathi. You are my everything.

PREVIOUS PUBLICATIONS

Parts of the ideas, figures and text presented in this thesis have appeared previously in the following peer reviewed workshop, conference, or journal publications. Use of such material is indictated in the footnotes.

Kahl, S. and S. Kopp (2015a). "Modeling a Social Brain for Interactive Agents: Integrating Mirroring and Mentalizing". In: *15th International Conference on Intelligent Virtual Agents*.

Kahl, S. and S. Kopp (2015b). "Towards a Model of the Interplay of Mentalizing and Mirroring in Embodied Communication". In: *EuroAsianPacific Joint Conference on Cognitive Science*.

Kahl, S. and S. Kopp (2016). "Communicative signaling and self-other distinction: Next steps for an embodied hierarchical model of dynamic social behavior and cognition". In: *13th Biannual Conference of the German Cognitive Science Society*.

Kahl, S. and S. Kopp (2017a). "Distinguishing minds in interaction: Modeling self-other distinction in the motor system". In: *3rd Workshop on Virtual Social Interaction*.

Kahl, S. and S. Kopp (2017b). "Self-other distinction in the motor system during social interaction: A computational model based on predictive processing". In: *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Kahl, S. and S. Kopp (2018). "A Predictive Processing Model of Perception and Action for Self-Other Distinction". In: *Frontiers in Psychology*.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

MNS	Human Mirror-Neuron System
MENT	Mentalizing Network
STS	superior temporal sulcus
TPJ	temporal parietal junction
IPL	inferior parietal lobule
IFG	inferior frontal gyrus
PMC	premotor cortex
TMS	transcranial magnetic stimulation
MT	Medial Temporal
mPFC	medial prefrontal cortex
pmPFC	posterior medial prefrontal cortex
HFA	high-functioning autism
SoA	Sense of Agency
PM	Person Model
HPBU	Hierarchical Predictive Belief Update
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
EBBU	Empirical Bayesian Belief Update
ToM	Theory of Mind
BDI	Belief, Desire, Intention
GAN	Generative Adversarial Networks
DMP	Dynamic Movement Primitives
EEG	Electroencephalography
fMRI	functional Magnetic Resonance Imaging

CONTENTS

ABSTRACT

Human communication often seems effortless. We tend to quickly have an idea of our interaction partner's intentions that enable us to predict their future behavior. How is such efficient communication possible which, despite uncertainty, allows us to quickly attribute beliefs to one another? Also, when and how are beliefs corrected if necessary? By investigating how action and perception influence and are influenced by prior beliefs during non-verbal communication, this work tackles the question of how and when the two subnetworks of the social brain interact.

A computational modeling approach is proposed, based on principles of predictive processing and active inference. The model's hierarchy consists of sensorimotor- and mentalizing levels. Their processes influence each other in a way that allows their embodied representations to be used efficiently. It is explored how uncertainty is handled in human communication, before examining the neuroscientific details of social cognition. Both inform the assumptions underlying the proposed model, which is evaluated in a number of simulations. These test the model's abilities to minimize uncertainty during action and perception, to differentiate between its own and other's actions, and also its ability to coordinate beliefs between multiple agents in a non-verbal communication game.

The simulations not only show that the proposed mechanisms quickly infer action intentions, able to influence future perception and action. Simulations also highlight the importance of weighting new evidence against prior beliefs, so that it is able to detect false beliefs and repair them during social interaction. The proposed computational model demonstrates a mechanistic account of the interplay within the social brain that allows for efficient non-verbal communication between similar agents, with implications for the notion of subjective direct access to other's minds.

1 INTRODUCTION

1.1 MOTIVATION

Social interaction can take many forms. Sometimes it is just a nod and a smile of a stranger in the streets, in the best of times you find yourself in a deep conversation with a good friend, and sometimes it even comes in the form of an unnecessarily hard volley during a tennis match. Most often, we find ourselves confronted with *uncertainty*, wondering about the meaning of other's behavior.

There are three questions which, when answered, can mitigate this uncertainty: One is, whether oneself is part of a social interaction: am I the target or addressee of a communicative act, e. g., is this person over there, who is gesturing or speaking, making this communicative act for me to understand? This question of being an addressee is trivial when I meet my neighbour alone on the street in front of my house, but think of a cocktail party, with a lot of people standing around in groups or alone. There, the communicative act of speaking or gesturing alone is not enough to make out the addressee.

Once we are sure that we are the addressee of a communicative act we face the second question: what are we communicating about? When thinking about the many ways to make yourself understood – verbally or non-verbally – and the myriad topics an interaction can be about, how is most social interaction so straight forward and efficient?

What we also need to answer is when we are done communicating, and whether we always need to be sure about our understanding? This third question is closely related to the second, in that the answer to it follows directly from inferring what the *communicative goal* is, e. g., in a process of *belief coordination* to reach a *shared understanding* (Clark, 1996; Clark and Schaefer, 1989; Traum and Allen, 1992). Once we think that we know what our goal is, we can track our progress toward it, but as you can imagine, knowing the goal is often not trivial. Now and then you find yourself thinking you reached your communicative goal, only to find that your interaction partner was trying to get a completely different message across. Human communication seems to be set up in a way that allows its practitioners to circumnavigate the pitfalls of handling this uncertainty, which can lead to *misunderstandings*. We seem to do this with ease, at least most of the time (Healey et al., 2018). But more often than not, we just assume our *mutual understanding* to be *good enough* without further investigation (Ferreira et al., 2002).

At first, one might think that the ease of understanding each other is based on properties of language in the form of verbal communication.

Yet, we will see that the same is true for non-verbal communication, which might even be closer to the origins of human communication, where in a form of *social motorics* the processes necessary for shared understanding have first come to be (Tomasello, 2008). Non-verbal communication takes many meaningful forms: from a body turned towards or eyes gazing at the addressee (Ciaramidaro et al., 2014), over the different kinds of gestures that represent something (Krauss et al., 2001; McNeill and Duncan, 2000), or joint actions on the world (Vesper and Richardson, 2014), down to small and meaningful deviations in all these non-verbal communicative acts, which can help to differentiate or make a point (Pezzulo et al., 2013).

The close relation between our body and cognition has been shown by studies of motor cognition (Decety and Sommerville, 2003; Gallagher, 2005), and gesture (Goldin-Meadow and Beilock, 2010). In that both, our own behavior and perceiving behavior of others, can influence our mental representations, and thus our understanding of each other. The theory of *embodied cognition* describes that cognition, as we possess it, requires a body, through which information can be acquired, or through which we engage with the environment and with our interaction partners in social situations (Wilson, 2002). In this perspective, most of what we can think about – or what we represent in our mind – is shaped by our experience of what is represented and filtered through our body and its sensory organs.

The human brain is specifically tuned to social interaction (Schilbach et al., 2008), perceiving other's behavior in the light of our own experience (Gallese et al., 1996; Keysers et al., 2004), and reasoning about our interaction partner's beliefs and desires (Schuwerk et al., 2014). Two functional subnetworks of the human brain have been identified to contribute: the mentalizing network (MENT), and the human mirror neuron system (MNS) (Van Overwalle, 2009). Yet, how the underlying brain processes work together to achieve this coordination of beliefs in social interaction, is still hardly understood and has been dubbed the "dark matter of social neuroscience" (Przyrembel et al., 2012).

One account of the processes of cortical function that takes uncertainty into account, is *predictive processing* (Clark, 2016; Friston and Kiebel, 2009). In this account, uncertainty about the information received about our environment through our senses is mitigated simply by correctly predicting the source of the uncertainty. This can be understood as a more general mechanistic property of efficient information processing systems that are in exchange with other social agents, and with their environment (Friston, 2013). Predictive processing is an account of passive information processing. To be able to influence the environment, or communicate with another social agent, it is extended to an account called *active inference*, where predictions can lead to action (Friston, 2011; Friston et al., 2010). This is similar to the *ideomotor principle* (Prinz, 1990).

This account of how uncertainty is handled in our brain could provide a framework for human communication. In this framework, the uncertainty one interlocutor has about her interaction partner's understanding needs to be reduced for this understanding to become shared.

1.2 OBJECTIVE

In this thesis I seek to develop a more holistic perspective to the general problem of how interaction partners can come to a shared understanding. Instead of an abstract discussion of human behavior in social interaction, here I create a *computational model* that aims to provide a mechanistic account and framework for handling uncertainty in social interaction.

To become more explicit: The purpose of the presented research is to find possible mechanistic and cognitive properties, as described in the literature, which underly the interplay of the processes within the human social brain. Further, the mechanistic properties should be able to reduce uncertainty between social agents during social interaction. In order to test the identified process hypotheses, they will be *computationally modeled* on the basis of predictive processing and active inference and *evaluated* in multi-agent simulations.

David Marr introduced a now widely applied *three-level* analysis to understanding information processing systems (Marr, 1982). To tackle the problems of reducing undertainty and coming to a shared understanding, *two* of Marr's levels of analysis are applied. The *computational* level is used to analyze the modeling problem and identify the necessary processes, so that in the next step, the functional accounts and computational modeling can be approached in the *algorithmic* level analyses.

The resulting identified functional accounts of the cognitive processes will be summarized in the modeling assumptions.

Two specific research questions will be addressed to come to a better understanding of the stated general problem:

- *How are action and perception informative in social situations?* In the context of this question we will visit and discuss aspects relevant for understanding social interaction in general. Then, we focus on belief coordination, non-verbal communication and the different findings from conversation analysis, hoping to uncover the core problem people face when they try to establish shared understanding with one or many interaction partners. Given the identified computational perspective on the problem, we will create a computational model based on necessary assumptions to first investigate the influence of action and perception during social situations. In order to handle uncertainty we will put the

model on the basis of predictive processing and active inference, investigating its suitability to handle the necessary processes.

- *Can active inference connect mentalizing and sensorimotor processing?* Following up on the first question, and with additional assumptions for the necessary interplay between sensorimotor and mentalizing processes, we will extend the modeling approach to cover mentalizing processes. We will investigate if active inference as an extension of predictive processing, allows to produce communicative signals for another social agent so that actual belief coordination based on perceived behavior and performed reciprocity can occur.

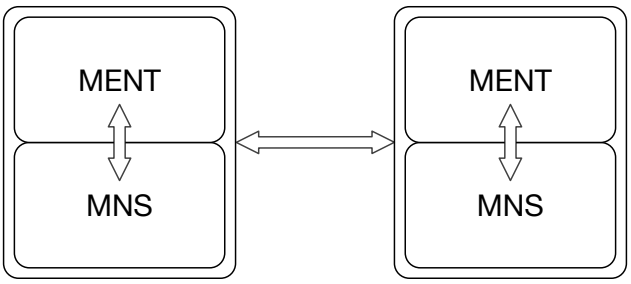


Figure 1.1: The basic interaction of only two social agents consists of an intra-personal interplay between the two processes within the social brain as well as an interplay between social agents.

This is the first time a computational cognitive model of the functional components of the social brain is created that can deal with the uncertainty inherent in social interaction. In addition, it is evaluated in simulations of multi-agent *reciprocal belief coordination*. A process of reaching a shared understanding between multiple interaction partners, based on a sharing of their respective beliefs (Clark, 1996). To achieve this, instances of the computational model are put into the social agents in an interactive environment (see fig. 1.1). This way the inter-personal interplay between multiple social agents can be modeled, with each agent having a model that incorporates the intra-personal interplay between the two processes within the social brain.

The model allows for belief coordination by applying different strategies to reach shared understanding through reciprocity. This is achieved, by either communicating a perceived belief, or by convincing the other agent of a belief by applying sensorimotor communication strategies, i. e., selecting an action that the interaction partner is more likely to correctly understand (Pezzulo et al., 2013).

At the core of the computational model lies an implementation of the predictive processing and active inference account which entails, as its primary goal, the need to minimize its *free energy* by predicting and correcting for the statistical irregularities in the signal (Friston, 2013). Free energy represents prediction errors of the system, as a

measure of uncertainty. The implementation of the model is a hybrid of a hierarchical Bayesian model and a linear dynamic model which, on each level, employs a variational belief update. This is designed as a form of continuous updating of beliefs based on the success of its prior predictions that minimize free energy. This continuous updating allows to accumulate evidence *online* during perception or action.

The model will be tested and trained to understand and produce non-verbal communication in the form of *handwriting of digits*, learned from a recorded corpus. Thus, the statistical irregularities that it has to deal with are deviations from previous movement directions. Such statistical irregularities are informative because they deviate from the previously applied predicting model, and lead to prediction errors, reflected by an increase in free energy. Each irregularity can have different reasons, which the model needs to infer, e. g., irregularities from deviating communication goals, intentions or action schemas.

The resulting representations will be the *embodied basis* that allows for belief coordination between the social agents.

The modeling and the evaluation results will demonstrate a mechanistic account of the interplay between mentalizing and sensorimotor processing – previously described as "the dark matter of social neuroscience" – with implications for the notion of subjective direct access to other's minds during social interaction.

1.3 OVERVIEW

Here, in the *first chapter*, we have now gotten a small motivatory introduction to the problems of uncertainty and shared understanding in social interaction. Also, I briefly summarized the actual objective of the following research and modeling, i. e., what the specific research questions involved are, and what the resulting modelings should encompass.

In the *second chapter*, I will go into much greater detail on the theoretical background. In the literature, we will see what it actually entails to be part of a social interaction, where context and our own prior information can play a vital role, when we align ourselves with our interaction partners in different aspects. I will also discuss the role of misunderstandings, and that they do not necessarily have to be fatal for communication, because we have different repair strategies. To further our understanding about the mechanisms underlying these central communication strategies, I will also look at the neuroscientific literature. There, we will get an idea of the complex interactions that make up the dynamic involvement of mirroring – or sensorimotor – and mentalizing activities within the social brain, which have been described as "the dark matter of social neuroscience".

In the *third chapter*, I will discuss additional background information relevant for the computational modeling. This covers introductions

to predictive processing, theory of mind, and motor control. Also, related work on other computational models of motor control and theory of mind will be discussed. And in addition, this chapter already covers the background and ideas for modeling representations under a predictive processing and active inference account of mentalizing.

In the *fourth chapter*, I will present the foundation of the computational cognitive model that contains many mechanistic properties – modeled on the basis of active inference – which are vital for handling uncertainty in exchange with the environment and during social interaction. Also in this chapter, the sensorimotor part of the modeling will be described.

In the *fifth chapter*, the established model, including its sensorimotor processes will be extended with a mentalizing part. This way, the two functional subnetworks of the social brain, along with their interaction, are put on the same modeling basis.

The *sixth chapter* contains the evaluation simulations that will test different assumptions and shine some light on the research questions I will formulate.

The simulation results, in context of the research questions, will be discussed in *chapter seven*, where we will also examine the general modeling approach and its implications.

In the *eigth chapter* I will conclude the presented work with a general summary, and a brief discussion of the contributions to different research fields. As a last word, we will also examine the limitations of the present work, and have an outlook at future possible work that would benefit from the presented computational model.

SOME REMARKS Before we come to the second chapter and dive into the theoretical background, there are some remarks:

- As you have probably already noticed, I here describe my research from a first-person perspective. Of course, not every result, and every line of text that was previously published, stems solely from me. Rather, it is an accumulation of ideas that went into the modeling assumptions, and into building the present model, and the evaluation simulations.

- I will repeatedly use “we”. This should be understood as standing for “me and the reader”.

We will now, as a first step, come to the necessary background on human communication from the perspective of linguistics, conversation analysis, and social neuroscience.

2

THEORETICAL BACKGROUND

In this chapter I will introduce the different aspects of *social interaction*, specifically *belief coordination* and the respective foundations and open questions, also from the perspective of the social cognitive neuroscience literature.

Social interaction: the communicative interaction between two or more agents with their own intentionality.

Belief coordination: the reciprocal back and forth resulting in mutual understanding.

In the first section (sec. 2.1) we will see that during everyday social interaction it is not only the simple act of speaking to our interaction partners that leads to successful communication but at most times we first establish a foundation upon which we are then able to efficiently communicate and coordinate our respective beliefs, namely a common ground. Following David Marr's three-level analysis to understanding information processing systems (Marr, 1982), in this section I will focus on the computational perspective on human communication and belief coordination. I will introduce the background of cooperative belief coordination and social resonance (sec. 2.1.1), with literature from conversation analysis and linguistics, which describe the important ingredients to form common ground during communication. These aspects will be discussed as a problem of reciprocity and uncertainty during communication, why most understanding may only be *good enough*, and one of establishing when communication was successful (sec. 2.1.2). Also, I will introduce several aspects of non-verbal communication (sec. 2.1.3), along with its importance in the evolution of human communication, because non-verbal communication will become a central aspect of the modeling approach that we will develop in the following chapters.

In the second section of this chapter (sec. 2.2) I will switch to a focus on the representation and algorithm perspective, in the discussion of the so-called *social brain*, and its two functional subnetworks: the mirror-neuron system (sec. 2.2.1) and the mentalizing system (sec. 2.2.2). This way, we will include our current mechanistic understanding of the one system, we know is able to communicate: the human brain. The two networks of the social brain are not independent of one another and their interplay and coordination will be the focus of the last section of this chapter (sec. 2.2.3). The exact process for their coordination (or their associated function) is in the locus of discussion for much of the literature on the foundations of belief coordination, and will be the vantage point from which I present my contribution.

Social brain: two partially overlapping networks of brain areas, functionally associated with social cognition.

2.1 BELIEF COORDINATION DURING SOCIAL INTERACTION

First, what is the foundation for communication that is established with our communication partner, so that we can efficiently communicate?

2.1.1 *From communicator to resonator*

Communication between two interaction partners is often way more than the suggested back and forth in encoding and decoding meaningful messages, which are being sent across channels, as described in the classical conduit metaphor of communication (Reddy, 1979). Many aspects of human communication are not covered by this metaphor. For example, simple encodings and decodings of messages cannot account for linguistic alignment (Pickering and Garrod, 2004), which describes several aspects of linguistic processes, in which communication partners align over time, or mimicry (Chartrand and Bargh, 1999) where gestures, posture, or choice of words are often unconsciously adopted by interaction partners to fascilitate likeability.

Social resonance: the collected coordination mechanisms at play during social interaction.

Kopp (2010) argues that the coordination mechanisms described here are always available to a certain extent during interaction and can be subsumed under the term *social resonance*. This is similar to the rapport (Tickle-Degnen and Rosenthal, 1990) that emerges between increasingly coordinated interaction partners when they experience a mutual attentiveness and coordination at different levels.

More generally, the more archetypal form of language use for communication has been described by Clark, in that *"[l]anguage use is really a form of joint action. A joint action is one that is carried out by an ensemble of people acting in coordination with each other"*(Clark, 1996, pp. 3). Thus, joint actions are a form of communication which entail the entrainment and coordination of overt behavior.

There are three types of coordination that operate on different time scales during social interaction, and they may to some extent be interdependent of each other (Kopp, 2010). In *behavioral coordination*, e. g., we see verbal and non-verbal adaptations of the body to an interaction partner. Then, there is belief coordination, which describes the back and forth of belief-representing communicative acts. Lastly, in *attitude coordination* the interaction partners let each other know their stance towards their joint goal of communicating. We will now discuss these in more detail, as they give a good overview of the dimensions of coordination that are part of human communication. Later, we will primarily focus on belief coordination as our modeling goal.

Behavior coordination: the overtly perceivable adaptations between interaction partners.

Attitude coordination: displays of cooperativeness, an exchange of platitudes to acquaint one with the other, or a simple exchange of gaze.

BEHAVIOR COORDINATION Mechanisms that are part of behavior coordination cover the overtly perceivable adaptations to an interaction partner. First of all, behavioral coordination manifests in linguistic

alignment, which has been observed in speech style, dialect, timing, prosody, intensity (e. g., Giles and Coupland, 1991) and speech rate (Street, 1984). But has also been observed in non-verbal behavioral adaptations during social interaction. There, this mimicry of body posture, facial expressions, speed and intensities of gestures and other mannerisms have first been subsumed under the term *congruence* (Kendon, 1973). Later it was referred to as the *chameleon effect* (Chartrand and Bargh, 1999), wich has been shown to enhance the smoothness of interactions while fostering liking among interaction partners. This form of mimicry is mostly unconscious, automatic, and has been suggested to be a form of social glue (Chartrand and Bargh, 1999; Lakin et al., 2003).

Congruence and Chameleon effect: the non-verbal behavioral adaptations between interaction partners.

Also, a temporal coordination between interaction partners or interactional synchrony has been reported, where listeners moved with the rhythms of a speaker's speech (Condon and Ogston, 1966). For example, a form of synchronized postural sway was reported by Shockley et al. (2003), where in a puzzle solving task the movements of participants, who were conversing with one another, synchronized more than in conditions in which they would converse with others. In a previous study it was suggested that it was the prosociality of a prime (e. g., suggestions that make the interaction partner more likeable), which increases mimicry, in contrast to antisocial primes that decreased mimicry (van Baaren et al., 2016). Following a different interpretetation of mimicry, Wang and Hamilton (2013) could show that for these primes to increase mimicry, they needed to be self-related and relevant for subsequent action.

Together, prosodic style, speech intensity, facial expressions and body posture are examples for verbal and non-verbal behavior coordination mechanisms that make you behave more similarly to your interaction partner and increase the likelihood of your *subsequent* communicative acts, and hence your communicative goal, to be understood. Thereby, these mechanisms help to tackle the second question faced during a social interaction, about answering what the communicative goal is.

BELIEF COORDINATION If the main function of communication is to make yourself understood, it comes down to what your intention to communicate is. A prime candidate for this is a belief you have, which can itself be about something that is in the world or an abstract concept you have in your mind. This belief has to be encoded and transmitted using whatever means you see fit to produce a communicative act, in order to display meaning for an interaction partner to understand and to respond appropriately.

Belief coordination is a highly dynamic and collaborative process of joint action that establishes a shared understanding, or *common ground* between interaction partners (Clark, 1996). The interaction partner's

Common ground: all shared knowledge, established prior or during the social interaction.

response plays an integral part in the belief coordination scheme, as it can demonstrate grounding acts, by which addressees can simply acknowledge a previously perceived communicative act (Traum and Allen, 1992).

Also, as we will discuss in more detail later, belief coordination can take multiple rounds of grounding acts from all interaction partners, who not only acknowledge but also demonstrate understanding, by means of reciprocating what is believed to have been understood (e. g., Swets and Ferreira, 2002). Grounding acts for feedback either acknowledge a communicative act or reciprocate a belief, and can inform interaction partners about their addressee's level of understanding and help them to formulate their next communicative acts (Clark and Brennan, 1991).

ATTITUDE COORDINATION Feedback, shared by interaction partners, is not only part of the belief coordination scheme, it also comes in the form of back-channel feedback, e. g., where an addressee displays a willingness to attend communicative acts (Allwood et al., 1992). This attitude coordination can entail displays of cooperativeness, an exchange of platitudes to acquaint one with the other, or a simple exchange of gaze.

This pertains to the first question you need to answer in any social interaction: am I the addressee of a communicative act, and hence am I part of a social interaction? Research by Garrod and Pickering (2004) suggests that during a conversation, sentence planning seems easier, and Swets and Ferreira (2002) show that the mere presence of an addressee seems to be a strong factor for the language production and understanding systems to make the most of the resources available, under the temporal constraints of online communication.

Social gaze: eye gaze not only displaying attentiveness, but also communicative intent.

As we will discuss in more detail later, *social gaze* especially has been shown to not only be a display of attentiveness, but it seems to be able to engage our full capabilities for social cognition to coordinate behaviors and beliefs in a gradual process of understanding (Myllyneva and Hietanen, 2015).

COMMON GROUND This gradual process of understanding each other would be immensely more difficult if you weren't able to presume your interaction partner to have a certain amount of prior knowledge. That is, common knowledge about your basic needs, like needing to eat or to sleep etc., or knowledge that you share about your upbringing as a human being. Malle (2001) discusses these aspects under the term causal history of reasons, as long as they play a role in forming an intention to act. These also entail the shared knowledge about having a body, which to a degree is mostly similar to that of your interaction partner, and for which you mostly share similar experiences of how it can be used to act upon and within the world. These experiences

also cover the interaction with other human beings, e. g., how you establish, develop, and end a conversation. Shared prior knowledge is either assumed or (partially) established during conversation, and is summarized under the term common ground (Brennan et al., 2010).

Now, instead of establishing every detail of your background and prior knowledge before each communicative act, you can just assume it to be part of the common ground you share with your interaction partner. This is the way in which common ground bootstraps communication (Clark, 1996).

Grounding: the process of establishing shared knowledge with your interaction partner.

New information can be introduced in a process of *grounding* which, when successful, updates common ground.

GROUNDING In the *grounding model* by Clark and Schaefer (1989) communication is a collaborative effort towards understanding to which all interaction partners contribute. This collaborative effort is described as a back and forth of making contributions, with identifyable phases. First, in the presentation phase, you make a contribution to the interaction, for your interaction partner to understand. Then, in the following acceptance phase, you observe whether your communicative act was properly understood.

Grounding model: a conceptualization of the process of grounding in human communication.

There are many ways for your interaction partner to provide this evidence of understanding. One would be to reciprocate what she has understood in her own words or gestures. Another is to simply acknowledge your contribution, by herself contributing a metacommunicative act – like nodding or smiling. Sometimes, it is even allowed for your interaction partner to just skip this acceptance, and continue with another presentation of her own to propell your dialog forward. This is allowed if her communicative act implies her understanding of your presentation, by that implying acceptance. To be exact, one should mention that even a display of acceptance could in itself be understood as a presentation.

What should your presentations and contributions consist of, i. e., when are they informative to the interaction? Grice defined a cooperative principle under which four categories of maxims for conversational contributions are specified, to *"[m]ake your conversational contribution such as is required, at the state at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"*(Grice, 1975, pp. 45-46):

MAXIM OF QUANTITY "relates to the quantity of information to be provided"

- make contributions as informative as required
- do not make contributions more informative than required

MAXIM OF QUALITY "Try to make your contribution one that is true"

- do not say what you believe to be false
- do not say that for which you lack adequate evidence

MAXIM OF RELATION "Be relevant"

- be relevant or make your contributions relate to the ongoing conversation

MAXIM OF MANNER "Be perspicuous"

- avoid obscurity of expression
- avoid ambiguity
- be brief
- be orderly

In this Gricean view communicative acts are not self-contained. Instead, meaning is coordinated through a grounding process, in which interaction partners seek and reciprocate evidence for understanding (Brennan et al., 2010), and by that update the common ground with newly established shared knowledge (Clark and Brennan, 1991). In this sense, grounding can also be understood as a process of joint hypothesis testing (Brennan, 1990), where temporary understandings – or hypotheses of meaning – are formed and tested constantly, by looking for misunderstandings or evidence for grounding.

Barr and Keysar (2002) argue that there is not much of a difference between processing in a dialog or monologue and that, e. g., language is first processed from an *"egocentric"* point of view. Only after that, an interaction partner's perspective is taken into account, in order to diagnose and correct coordination problems, e. g., in the form of just previously established word meanings.

As a more effortless approach to establishing a shared understanding the *interactive alignment model* argues that processing of communicative acts in dialog is different than in monologue, because in the former, the comprehension and reponse planning processes need to be active at the same time (Pickering and Garrod, 2004). This is because they assume that production and comprehension processes are based on the same linguistic representations which, when activated during listening to your interaction partner, can become *aligned*: similar mental representations are active in all interaction partners. Further, aligned representations can prime future processes of comprehension and production. For example, during listening, an interaction partner's language primes your response planning to make use of similar

grammatical structures, or words. The authors also assume that the activation of representations can activate others (also at different levels of processing). Other than the grounding model by Clark, no explicit common ground needs to be established. Instead, the interactive alignment model proposes that aligned representations form an implicit common ground.

Adaptations (e. g., Kopp, 2010) happen at different levels and can happen anytime. Clark and Krych (2004) show this in studies of cooperative building of Lego models. They find that the interaction partners' behavior is highly coordinated, rapidly adapting to each other. Making several observations with implications for models of speaking, they state that the most basic implication is that speakers make use of their processing capabilities differently during dialogue, then when they are alone. They observe that interaction partners update common ground *all the time*, not just after each turn (as in the grounding model). Also, speech is constructed jointly by all interaction partners, evidence for understanding is given as soon as possible (using all available communication channels), and speakers early on during speech production plan to have to make repairs later on.

This is not a full list of accounts of grounding, but in most, the information provided by an interaction partner is taken into account to either prime yourself for a specific understanding or be sure about the other's correct understanding.

SOCIAL RESONANCE When we want to define what a successful social interaction looks like, we need to identify the process of reaching understanding. A prerequisite for this process is that it requires a shared *communicative intentionality*, which has be conveyed in behavioral and attitude coordination (Tomasello, 2008). For successful communication common ground needs to be established between interaction partners, which is achieved in a dynamic grounding process, in which communicating agents reciprocally reveal and coordinate their beliefs about each other as well as the state of their interaction (Clark and Brennan, 1991). The different coordination mechanisms are not separate, but go hand in hand to establish familiarity, trust and rapport, forming a foundation for social resonance, upon which communication can be successful, and thus, knowledge can be shared (Kopp, 2010).

Communicative intentionality: the perceived willingness to engage in communication and perception of another's mind to have beliefs on its own.

The process of establishing the shared knowledge during social interaction is not an easy one and communication is not always successful, since the grounding process described here is prone to errors. For example, Clark and Schaefer (1989) suggest that the question of when a new aspect becomes shared understanding depends on a so-called *grounding criterion*. It dynamically adjusts to the need for understanding in the current context, either allowing it to be only

shallow, or *good enough*, or expects it to reach a more solid level of understanding.

2.1.2 *Good enough understanding in social interaction*

In psycholinguistic investigations on sentence understanding, we can find relevant pointers that hint at the foundations underlying general belief coordination and grounding in conversation.

First of all, linguistic interpretation is very fragile and needs immediate reinforcement by context, experience or feedback from an interaction partner (Sachs, 1967). Also, during the comprehension process of sentences with difficult syntactic structures, like garden-path or implausible passive sentences, its intial interpretation can be unstable. A classical example for a garden-path sentence is *"A horse raced pass the barn fell"*, and for you to interpret it correctly you have to revise your initial interpretation and chose another one instead. It was shown that when asked about specifics about such sentences, your first interpretation *lingers* and can influence your subsequent sentence understanding (Christianson et al., 2001).

When it comes to implausible passive sentences, Ferreira and Stacey (2000) showed that people seem to apply world knowledge to derive who is doing what to whom, in a form of fast and frugal heuristics, rather than proper syntactic algorithms. For example, in contrast to the sentence *"The man bit the dog"*, a more unpredictable sentence like *"The dog was bitten by the man"* is often misinterpreted when asked about who did what to whom. Context helps to stabilize possible interpretations of such communicative acts that are hard to interpet, and they argue that such syntactic instabilities are usually no problem during normal conversation, because the communicative context would support the interpretation.

Ferreira et al. (2002) suggest, that for an interpretation to become more likely, enough information to support it must be collected. Otherwise, only a *good-enough* understanding of a sentence is derived, i. e., a possible understanding which only superficially satisfies the given information constraints, instead of having done a time-consuming and in-depth evaluation of all possible understandings (see also Simon, 1955 on rational choice and satisficing humans).

One could argue for the computational advantages of a system that initially underspecifies, and fills in information, as the details become relevant and available (Sanford and Sturt, 2002). For example, Pickering and Frisson (2001) argue that delays in disambiguation allow for multiple meanings to become probable from context, while an underspecified meaning is active. This also may reduce frequency effects, where simply the most often seen use of the sentence becomes selected.

This approach suggests a necessary underspecification, because of a capacity limit and goal directedness of the comprehension system. That outside of a linguistics laboratory, people are seldomly asked detailed questions about their understanding, may be another hint at the incomplete and shallow representations, which we derive.

It has also been argued that a cognitive system that considers all relevant information to arrive at a decision is biologically unrealistic, due to the limits of resources humans must obey (Gigerenzer et al., 1999). Limited resources, such as time pressure during a conversation, or the limited capacity of working memory, may be key factors for the incremental nature of conversation.

For example, Swets and Ferreira (2002) found that time pressure had the effect of making speech more incremental. Similarly, Swets et al. (2013) used time pressure in an interactive tangram description and matching task, to investigate its effects on speech production. Participants either had the matcher role or the director's role. The matcher had to repeat the director's description and match the tangram figure, while the director was under time pressure to start describing. They found that it was not time pressure that had any significant effect, but the mere presence of the interaction partner made the director's utterances more descriptive.

MISUNDERSTANDINGS Swets and Ferreira (2002) state that the interaction partner's mere *responding* is what determines the success of the joint activity, as it is the only overt and interpretable signal of the interaction, and we seem to rely on the slightest cues that allow us to interpret their understanding in their favor.

A recent radical interpretation of *embodied cognition* by Wilson and Golonka (2013) makes a similar assumption for the depth of understanding during social interaction. This assumption entails that the information conveyed is not as important for the successful discourse, as is the mere perception of the addressee's understanding. They argue that in the subjective experience of an organism, there is no difference between the meaning of perceptual and of linguistic information. This means that if you can *correctly* use linguistic information to meet the expectations derived from your interaction partner's intended meaning, you have also demonstrated that you know the meaning of that information, even if you actually lack proper understanding. What is relevant to this argument is that the mere *assumption* of understanding can be deemed *good-enough*. In the incremental process of belief coordination, this good-enough approach relieves us of the burden of needing to make sure that every said word was properly understood, and instead focus on the relevant parts.

Relevant in a conversation are not only the beliefs that we intent to communicate, but in order to create a stable context for linguistic interpration, and a common ground, we also need to focus on misun-

derstandings. Misunderstandings are detected during interaction, or after a supposedly successful belief coordination attempt. An example for this is when you learn the vocabulary of a new language, but then during class, you are not entirely attentive, so you end up learning the wrong meaning. In the best case, your misunderstanding will just lead to some laughs, when you then try out your new learned language skills, after which you will be corrected, and you learn the correct meaning. In other cases, your misunderstanding will just go unnoticed.

Misunderstandings are not fatal for social interactions but can be repaired. *Repairs* describe *"the methodical practices provided in the organization of talk-in-interaction for dealing with problems or troubles in speaking, hearing, or understanding the talk"* (Schegloff, 1987, pp. 110). There is a long standing research tradition on dealing with difficulties with mutual understanding (Sacks et al., 1978; Schegloff, 1987, 1995), and several models of semantic coordination habe been developed, driven by clarification requests, detecting communication errors and to initiate repairs (Eshghi et al., 2015). All these approaches to dealing with problems of understanding, speaking or hearing of a talk, take into account the (mis-)understood aspect, evaluate it in the context of the ongoing interaction, and can result in the request for a clarification. For example: reciprocity, as the own production of the aspect previously understood, is one possible way to implicitly request a clarification.

Running repairs hypothesis: the idea that repairs during communication are not the exception, but the norm.

In recent work, Healey et al. (2018) discuss their *running repairs hypothesis*, which has this exact focus: reciprocity not only is a tool of belief coordination, but also helps to collect context information. Thus, it stabilizes a (linguistic) interpretation by allowing parts of a communicative act to carry uncertainty, but still be important for disambiguation.

This is a vital point, which we will come back to later, when the model of belief coordination will be described.

BELIEFS ABOUT SELF AND OTHER GUIDE COMMUNICATIVE ACTS

A very important aspect in the conversation about how shared understanding is reached, is how our beliefs are formed, represented and utilized.

Meta-cognition: the process of thinking about thinking, which may even allow to steer our thoughts.

This pertains to the role of *meta-cognition*, or thinking about thinking. An aspect of cognition that allows us to revise and (to a degree) steer our thoughts. During social interaction such meta-cognition is often described as theory of mind, where we form beliefs about the contents of our interaction partner's and our own beliefs, desires and intentions (Premack and Woodruff, 1978; Rao and Georgeff, 1995).

COMMUNICATIVE SIGNALING

One way in which meta-cognition during social interaction can utilize perceived beliefs about our inter-

action partner, is by making subsequent communicative acts directly dependent on those beliefs, e. g., to confirm or disconfirm what is believed to have been perceived.

This is similar to the running repairs hypothesis, where parts of a communicative act needs to be explained in more detail, or has been detected as being misunderstood (Healey et al., 2018). A false belief would be used as a negative exemplar, from which your next communicative act should be clearly differentiated.

Since this is very abstract, imagine yourself describing to a friend your way to work, using co-speech gestures. After the first description your friend stops you, so that she can repeat what she has understood. At a crucial point, where you have to take the correct exit at a tricky roundabout, she makes a mistake. In response, you explain this part again, but this time you make sure that she understands to take the correct exit by explicitly drawing the circular way you take through the roundabout, up until the correct exit. This time, she understood correctly.

Through your explicit drawing gesture, you have *"signaled"* your intended action. Such strategies to signal intent are often embedded in pragmatic actions as understandable kinematic signatures. The literature also refers to this in the non-verbal communication domain as *sensorimotor communication*.

Sensorimotor communication: taking someone else's perspective into account, to make action kinematics easier to disambiguate.

It is a focus of interest in recent years, e. g., in the literature on so-called *joint action*. There, interaction partners infer each other's intentions and goals through their respective actions by a process of tight dynamic coupling in joint behavior tasks. These tasks can take the form of simple interactive settings with two participants. In a study of synchronous tapping on specific targets, interaction partners informed each other of the target, by exaggerating the amplitude of their trajectories (Vesper and Richardson, 2014). Konvalinka et al. (2010) found that participants were good at synchronizing in a joint tapping task with an interaction partner that was both, tapping regularly and in a responsive manner. They were also synchronizing with an irregularly tapping and responsive other. Responsiveness here means that participants adapted to one-another, using slight variations to brought their tapping frequency closer to their interaction partner's frequency. At the same time, this coordination could not be established with an unresponsive computer, that nonetheless was predictable. It seems that the mutual predictability and reponsive adaptation on a millisecond timescale is important, but could not be found in interaction with the unresponsive (and non-adapting) computer.

A major result from the research on joint action is that participants in such tasks tried to reduce their action variability in order to become more predictable. Another major finding is that to increase the predictability of their actions, people often strategically change action

kinematics, to make them easier to disambiguate. For example, emphasizing the amplitude in the movement between taps in an interactive tapping task, to highlight the time between taps. Pezzulo et al. (2013) modeled the dynamic interachange of gesture trajectories during joint action. There, gestures had to be optimized toward maximum discriminability with respect to other, equally possible, action trajectories. They found that deviations from an action's optimal trajectory could be parameterized in such a way that its original pragmatic goal is preserved while making its kinematics informative for another action. They describe sensorimotor communication as an intentional strategy during joint action that supports social interactions by making action kinematics easier to disambiguate. The authors assume a predictive account of action understanding and hence argue that such signaling must have the goal to increase the predictability of the actor's intent.

More recently, Vesper et al. (2016) found that there is a trade-off between trying to coordinate through sensorimotor communication and reducing variability to be better predictable. In contrast to information transfer, in true communication, signs are selected for a communicative purpose from a communicative intention towards an interaction partner (Clark, 1996). Additionally, sensorimotor communication has to be distinguished from conventionalized forms of communication, i. e., learned code such as spoken language or sign language, in that it has a specialized aspect to it. For example, it acts like a deictic gesture (like pointing towards something), rather than an iconic gesture (convey conventionalized meaning that is also present in speech). Sensorimotor communication can be described as a combination of a pragmatic communicative goal with an additional specialized signal, e. g., carrying a table together, where by applying slight pressure on one end, you signal in what direction you want the table to be carried.

This comes as another aspect to belief coordination and reciprocity, as it allows to select communicative acts, and act, by taking specific aspects of a perceived understanding of your interaction partner into account. Beyond sensorimotor communication, this can also be imagined in different modalities, be it the pronounciation of words for the goal of clarity in a noisy environment, or the pragmatic choice of words that allows for improved understanding in the listener, and the example from the described literature, that is the alteration of a kinematic trajectory of an action for signaling purposes.

Since the described strategies for sensorimotor communication can be a valuable aspect of the process of belief coordination, wich can best be observed during non-verbal interaction, we will now discuss the application and role of gestures for human communication.

2.1.3 *Non-verbal communication*

A specific focus of the present research is non-verbal communication, such as writing, gestures or social gaze. This is due to its unique importance in the evolution of human communication. In that, Michael Tomasello sees a strong case for the thesis that non-verbal communication like pointing, gesture, and gaze are a *"primordial form of uniquely human communication"* (Tomasello, 2008, pp. 3) and are the foundation, in which later verbal skills of communicating are rooted.

Pointing gestures seem to be based in cooperative communication, since in itself such a gesture would mean nothing. It is only in the situation, with shared situational context and joint attention between interaction partners that a mere pointing carries information that is shared. Tomasello argues that a complex psychological *"infrastructure"* (Tomasello, 2008, ch. 3) is at work to achieve this, which at some point in human evolution granted individuals an adaptive advantage to engage in joint intentions, joint attention, and cooperative motives. He constructs and describes the components of the necessary infrastructure, which he calls the cooperation model.

When looking at the ontogenetic development of humans, one finds that already one-year-old children begin to point and pantomime. These are gestures that require the child to understand their interaction partners to be intentional agents, with whom such joint attention is possible. Tomasello observes that language use in children starts shortly after gesturing, and that their ability to perceive someone's intention is crucial for using communicative gesture, and later words. Importantly, he argues that it is not the learning of the code of language in which language ability originates, but a non-verbal (or non-linguistic) infrastructure of intentional understanding and common ground.

In contrast to the seemingly intuitive understanding of intentionality, which even young children have, Tomasello describes the social behavior of great apes, where most primate communication is studied. There, he describes a mixture of genetically fixed and inflexible gestures, and in addition a subset of gestures that seem to be individually learned, and should be called intentional signals. While apes make great use of these intentional gestures, flexibly display social intentions and seem to understand the intentional actions of other apes, they do not account for shared intentionality. That is, they do not relate a gesture to the inferred intention of an interaction partner. By that, they neither establish, nor share common ground explicitly.

Since we can still find cooperation and social interaction in great apes, these biological adaptations are likely to provide an evolutionary advantage. And while we can find individual and flexibly-used intentional gestures, it was found that these do not spread through a community, like human conventional signs would do. Taken together,

Tomasello argues against what still too many linguists assume, i. e., some version of an innate universal grammar. He makes a solid case, arguing for a more empirically grounded approach, from which it was observed that very important prerequisites are missing from our closest evoluationary relatives. For example, the ability for a human-like establishment of common ground, or the process of belief coordination. Equipped with this, even the richest intentions can be conveyed by simple gestures, because of our uniquely human ability to share common ground with our interaction partners. So, while probably not possible without the prerequisite of understanding intentionality, human language is culturally constructed, and not biologically inherited. Gestures, nonetheless, form the evolutionary basis of language, and the main difference may be found in the "infrastructure" underlying human-like establishing of common ground, and the process of belief coordination.

WRITING AND GESTURES Many gestures occur only during speech-use, which makes them co-speech gestures. These can be beat gestures, which carry emphasis and are produced to emphasize specific prosody during speech, or iconic gestures, which are conventionalized gestures, heavy in semantic content that can complement the meaning conveyed in speech. Symbolic gestures are conventionalized, can be culture specific, can replace words, and other than iconic gestures can occur without speech (e. g., a thumbs-up gesture, or waving for saying "hello") (Krauss et al., 2001). Also deictic gestures, like pointing to something, play an important role in communication of great apes (McNeill and Duncan, 2000).

Usually, gestures are imagistic and sporadic when they are developed while accompanying speech (Goldin-Meadow, 2006). Also, they often convey information not carried by speech, so the burden of communication is shared between speech and gesture. In contrast, gestures can develop a language-like form when the whole of communication depends on gesture alone, like in sign language. These phenomena were studied by Goldin-Meadow (2006), who found that deaf children without exposure to conventional signed language spontaneously invent gestures, to communicate in rich and language-like ways. Research on sign language shows that gesture can be as semantically rich as verbal communication.

Similar to iconic gestures, writing is a form of non-verbal communication that occurs without speech. It is a specific form of speech-related gesture that interacts with the language processing systems responsible for speech and gesture. This has been shown in stuttering subjects, who were able to perform motor tasks, but not writing during moments of stuttering (Mayberry et al., 1998). Writing has also been described as a form of gesturing in a philosophical essay by Vilém Flusser: *"To write is to in-scribe, to penetrate a surface, and a written text is*

an inscription, although as a matter of fact it is in the vast majority of cases an onscription. Therefore to write is not to form, but to in-form, and a text is not a formation, but an in-formation. I believe that we have to start from this fact if we want to understand the gesture of writing: it is a penetrating gesture that informs a surface." (Roth, 2012, pp. 26).

Gesturing and speech can occur simultaneously, but also the occurrence of gesture can happen automatically, without conscious action. In several studies, participants' behavior was analyzed while talking on the telephone. It was found that regularly *co-speech* gestures are used, as if the interaction partner was visible (Bavelas et al., 2008). Also, the production of co-speech gestures has been found to be modulated by interaction partner visibility and intentionality of the interaction partner (e. g., in one condition Bavelas et al. (2008) had participants talk to a tape recorder). Another study found a difference in the frequency of co-speech gestures when visibility between interaction partners could be blocked by a screen (Alibali et al., 2001), and although there was a difference, gesturing was never absent. This points to an interpretation of the representations of speech and gesture that connects them intimately.

As has also been shown by McNeill and Duncan (2000), there is a strong overlap between brain areas that are active during the perception and recognition of speech and gesture. This points to the idea that both speech and gesture are strongly intertwined. McNeill also offers the view that gestures are elements in the cognitive process itself, in that *"the actual motion of the gesture itself, is a dimension of thinking"* (McNeill, 2008, pp. 98). By that, so he argues, gesture is part of a loop of self-directed speech – similar to writing – that can help in your own thinking process.

EMBODIED COGNITION The studies that describe the intertwined nature of action and perception argue along the same line as the theory of *embodied cognition*. Embodied cognition describes that cognition, as we possess it, requires a body, through which information can be acquired, or through which we engage with the environment and with our interaction partners in social situations. In this perspective, most of what we can think about – or what we represent in our mind – is shaped by our experience of what is represented and filtered through our body and its sensory organs. Wilson (2002) evaluates the main claims of embodied cognition, concluding that (adapted from her paper, pp. 626): (1) cognition is situated: it takes place in the context of the environment, involving perception and action; (2) cognition is time pressured: its function is optimized for real-time interaction; (3) off-loading of cognitive work on the environment: we make the world hold or manipulate information, retrieving it if needed; (4) cognition is for action: the function of the mind is to guide action; (5)

Embodied cognition: embodied cognition describes that cognition, as we possess it, requires a body, which shapes our perception, cognition, and action.

off-line cognition is body based: even uncoupled from it, cognition is grounded in mechanisms evolved in interaction with the environment.

Also, mental representations associated to a gesture have been shown to be influenced by performing or seeing that gesture (Goldin-Meadow and Beilock, 2010). Also, Gallagher writes that *"even if we are not explicitly aware of our gestures, and even in circumstances where they contribute nothing to the communicative process, they may contribute implicitly to the shaping of our cognition"* (Gallagher, 2005, pp. 121). He suggests that gesture is a means by which thought is accomplished, and at the same time, an aspect of the thinking itself. Similarly, not only own actions influence these mental representations, but also those performed by others. Decety and Sommerville (2003) argue that this close connection underlies *motor cognition* in general, as a means to think about and handle our own and other's actions. Further they argue that this encompasses all levels of cognitive processing. Those involved in the generation of our own action, and also in the prediction and understanding of other's actions.

2.1.4 *From behavior to neural processes*

I have chosen to focus on non-verbal communication, because of the evidence for it being an ontogenetic and phylogenetic precursor for verbal communication. Specifically, writing with a pen on a piece of paper can be described as a highly conventionalized form of drawing gestures with widely-recognized meanings, and will be used in this work as the domain for communication.

Also, the evidence for an embodied nature of cognition sheds light on the intertwined nature of motor cognition and social cognition. Understanding the neural correlates underlying these processes could give us more than a hint at how the process of blief coordination works for communication in general.

From the Marr'ian computional level of analysis of looking at human communication, we will now switch to one of representation and algorithm in the discussion of neural correlates of social interaction in the so-called social brain – two partially overlapping functional networks in the human brain, most active during social interaction.

2.2 THE SOCIAL BRAIN

With the goal of better understanding social interaction in humans, it is not only the dynamic interplay *between* interaction partners that should be a research focus. It is equally important to understand the function of communicative signals, like language and gestures, as it is important to shed light on the neural correlates for the processes underlying the understanding and production of such signals in the individual brain, in an interactive context. A first correlate for the

process of *understanding* an interaction partner's action, is the Human Mirror-Neuron System (MNS). In addition to action understanding, people infer the intentions behind other's behavior. This discrete ability to predict and interpret *social behavior* is often referred to as *mentalizing* or *Theory of Mind (ToM)*. Brain regions associated with mentalizing are part of the Mentalizing Network (MENT), as part of the social brain.

ToM: Theory of Mind is the cognitive ability to reason about another person's mental states.

2.2.1 *Human mirror-neuron system (MNS)*

MNS: the human mirror-neuron system is a functional subnetwork of the social brain, which is active both, during perception and production of action.

Back in the 90s, di Pellegrino et al. (1992) first reported an interesting response they had observed in monkeys. Recording the electrical activity of neurons in Brodmann area 6 (F5 in monkeys), where neurons have been associated with grasping behavior, they found that the same neurons that are active during the performance of grasping, are also active during the mere perception of somebody else performing a grasp. In another paper, the role of mirror neurons in action recognition is discussed, while already positing their existence in humans (Gallese et al., 1996).

Human mirror neurons were eventually detected *in vivo* (Mukamel et al., 2010), but instead of isolated frontal and premotor regions, as in monkeys, mirroring activity was recorded in the human homolog as well but also in supplementary motor areas, hippocampus, and their environment. These *in vivo* extracellular electrode recordings found many different areas to be exhibiting mirroring activity, but mainly such activity was found in sensorimotor convergence zones.

In addition to the observation that it is a multitude of systems that perform sensorimotor mirroring activity, there soon was evidence for multisensory (also covering other senses than sight and proprioception), and socially relevant mirroring activity, e. g., a pain-response to observed cues for anticipated pain (Hutchison et al., 1999). There is even evidence for an involvement of the insula-striatal system in the recognition of disgust in others (Calder et al., 2000), and activity has been found in the secondary somatosensory cortex while the human participant was touched as well as when the participant observed somebody or something else being touched (Keysers et al., 2004). Also, not only actually perceived actions do trigger mirroring activity. As it was found by Grossman and Blake (2001), even *imagined* biological motion was sufficient to trigger activity in the *superior temporal sulcus (STS)*, as part of the mirror-neuron system.

STS: a brain area that has been associated with processing visual information of behavior, without differentiating between self and an interaction partner.

BRAIN AREAS OF THE MNS Because of the widespread areas that respond to observations and performances of actions, in human imaging studies it has not been straight-forward to pinpoint exactly which areas are involved in the mirror-neuron system in humans.

I will now briefly summarize the cortical areas of the MNS (visualized in fig. 2.1), adapted from a meta-analysis, which collected over

200 functional Magnetic Resonance Imaging (fMRI) studies and identified major functional brain areas that are involved in social cognition (Van Overwalle, 2009): One major area that has been strongly associated with processing visual information of social settings, is area STS, which processes visual information of behavior, without differentiating between the self or an interaction partner. In a supposed hierarchy of information processing, the next processing step can be associated with the *inferior parietal lobule (IPL)*, which is strongly overlapping with the *temporal parietal junction (TPJ)*. Both areas have been associated with inferring the intention of a movement, with a specificity to social information, and the function of identifying the agent of a social action as the self or distinct from that. Further up the hierarchy, information is send to the *premotor cortex (PMC)*, an area involved in high-level processing of own actions, i. e., it has been suggested, that in social cognition the PMC handles a comparison of the visually inferred behavior of self or an interaction partner, with own action schemas and their goals. This match of action schemas, along with information about the action's future path, is passed back to the IPL/TPJ area, so *"In a sense, the IPL "sees" the intentions behind other's actions by "simulating" or "matching" the actions of others in a shared representation"* (Van Overwalle, 2009, pp. 831).

TPJ: a brain area that has been associated with inferring the intention of a movement, and identifying the agent of an action.

PMC: the premotor cortex has been associated with visually inferring or comparing own action goals with that of others.

Throughout the literature on mirror neurons it is often discussed how sensorimotor mirror neuron activity underlies human social life, or whether a general mirroring mechanism underlying all mirroring-like properties, found throughout the human brain, may be responsible. In one article Gallese et al. (2004) discuss a potential unifying basis of social cognition, covering not only actions but also emotions of others. They end their discussion saying that potentially *"[s]ocial cognition is not only thinking about the contents of someone else's mind. Our brains, and those of other primates, appear to have developed a basic functional mechanism, a mirror mechanism, which gives us an experiential insight into other minds."* (Gallese et al., 2004, pp. 401).

Given the very specific kinematic patterns that are influenced by specific prior intentions (e. g., like avoiding to run into one another when we walk opposite directions on the street), we need to assume that it makes sense for humans to be able to pick up very minute kinematic patterns, and infer these specific intentions early on.

ORIGINS OF MNS Whether such a mechanism for such "insight" is *innate* to us, i. e., given to us from birth, or whether it *develops* – and to what extent – is hotly debated.

On one hand, there is the hypothesis for mirror-neurons to have developed as an adaptation to evolutionary pressure, to understand what other individuals were doing (this is implicit, e. g., in Rizzolatti and Arbib, 1998) in ever growing social structures (cf. Dunbar, 1998) – be it in apes or their human relatives. In this view of mirror-neurons,

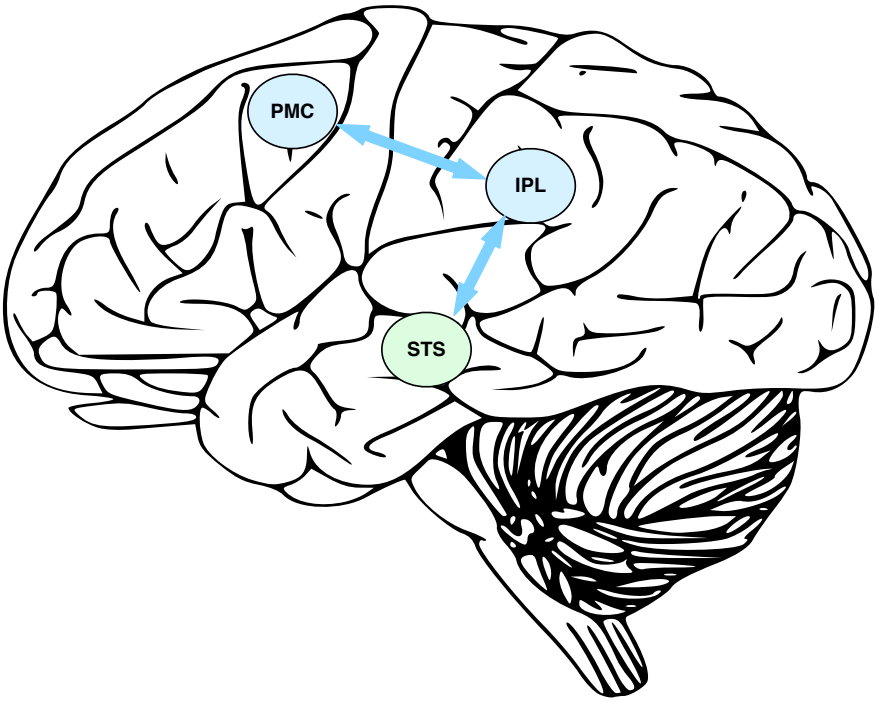


Figure 2.1: The human mirror-neuron system consists of several functional brain areas that have been identified in a meta-analysis of over 200 fMRI studies (Van Overwalle, 2009). Specifically, area IPL has repeatedly been shown to be active in motor intention inference. Also, the PMC area has been associated with monitoring own behavior, and even with the ability to differentiate own from other's behavior, and focusing attention. STS on the other hand does not distinguish between self and other behavior, and can be seen as a first hub for visual information being fed into the social brain.

sensory or motor experience is believed to facilitate their development, but their function to match observed and executed actions is genetically inherited.

Association hypothesis: the hypothesis that mirror-neurons develop through an association process between sensory and motor activity.

A different hypothesis for the development of mirror-neurons is the *association hypothesis*. The hypothesis is that in brain areas, which allow for neuron activity, in response to a sensory modality and in response to motor activity, an association between these activations in close temporal proximity is formed. This may result in learned sensorimotor associations, in response to the perception and production of the underlying cause of the neuronal activity – similar to Pavlovian conditioning (e. g., Heyes, 2001, and for a review see Heyes, 2009).

In support of the association hypothesis, it was found that premotor transcranial magnetic stimulation (TMS) stimulation enhanced mirror-neuron motor facilitation as well as the effect of prior counter-mirror training (Catmur et al., 2011). Counter-mirroring describes the neuronal activity in response to a sensory stimulus, where the motor neuron activity does not correspond to the associated sensory

effect. During training, participants perform a movement while they observe a movement that involves some other muscle, so that during observation you can find motor neuron activity for that other muscle.

The association hypothesis does not preclude any brain areas from developing mirroring activity, and thus may better be able to shed light on the development of mirror-neuron activity in humans, which is not only being found in classical motor areas.

What is important for our later approach to modeling, is that context strongly influences the understanding and production of an action. For example, Georgiou et al. (2007) showed how a specific prior intention (*"individual vs social, cooperative vs competitive"* Georgiou et al., 2007, pp. 432) leads to very specific kinematic patterns in the same action. Also during perception, context was found to strongly influence what information is most useful to the participant (Streuber et al., 2011), and how a scene is interpreted (de la Rosa et al., 2014).

A point not often made, is that there markedly seems to be a difference between human and monkey mirroring activity. In contrast to monkey's mirror-neurons only firing on the transient activity, i. e., goal-directed action, there is evidence for human mirror-neuron activity to also occur during intransitive actions (Bertenthal et al., 2006; Press et al., 2008).

COMMON REPRESENTATIONS FOR ACTION AND PERCEPTION Much of the literature argues that the capabilities to perceive and make sense of our interaction partner's actions are basic prerequisites for being socially resonant. and are rooted in a sensorimotor basis, like the human mirror-neuron system, which has repeatedly been found active for social behavior. Congruent with the discussed capabilities, associated with the human mirror-neuron system, Hommel et al. (2001) proposed a *theory of event coding*, in which observed (action-) events, and own planned actions, are encoded in a common representational medium as bundles of so-called feature codes. On the same line of argument, the *common-coding* hypothesis was brought forward (Prinz, 1990, 1997).

Common coding: the hypothesis that representations used for understanding an action and planning an action are encoded in a common representational medium.

These ideas and those highlighted earlier all provide a theoretical basis for the embodied processes in humans, underlying action understanding, learning and planned action execution, along with effects of anticipation and priming. One big question in the discussion on the interplay within the social brain touches the way in which our mind interacts with, and understands, the environment by means of the body.

As already described, this has come to be called embodied cognition, which *"[. . .] stresses that perception and action are directly relevant for our thinking, and it is a mistake to regard them as separate."*(Willems and Francken, 2012, pp. 1). Several claims in the embodied cognition literature are relevant to us, and have been reviewed by Wilson (2002). For example, our cognition takes place in the context of a task at hand

in an environment that we perceive, i. e., it is *situated*. So in some sense the environment becomes a part of our cognition.

A claim in the literature that has also gained much interest, is that cognition is *for action*, which was rooted in the early works on the MNS (as reviewed above), about motor neuron activity in response to the right visual stimulus. That being said, it has also been argued that a purely sensorimotor basis may not be enough for social behavior. E. g., Jacob and Jeannerod (2005) discuss evidence for differential activity in non-human primates, for actions directed towards conspecifics that led to purely perceptual neuronal responses, without motor activity. They argue that the motor system alone might not be well equipped to elicit appropriate responses to social behavior.

The evident human ability to predict and interpret social behavior is often referred to as mentalizing, and related tasks often found activity in the second subnetwork of the social brain: the mentalizing network.

2.2.2 *Mentalizing network (MENT)*

MENT: the mentalizing network is a functional subnetwork of the social brain, which is active during interpretation of an interaction partner's social behavior.

One of the most influential, and widely known works that show the acute human sensitivity to social interpretations, are the drawings and animations from Heider and Simmel (1944). They trigger associations with social situations, and emotional responses to the shown behavior. The inference behind these associations require a form of reasoning about the potential latent mental states, causing the behavior. In a series of animations, simple geometrical figures, through their movement alone, create the impression of intentional behavior (for a small example, see fig. 2.2).

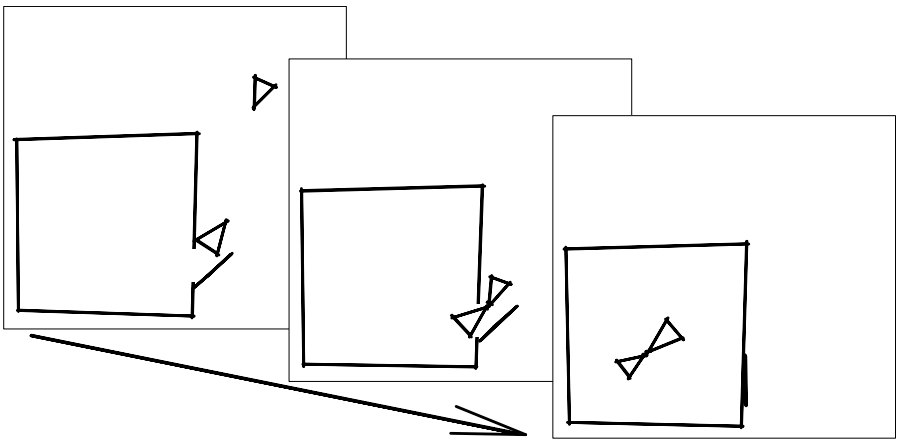


Figure 2.2: In their paper on "An experimental study of apparent behavior" Heider and Simmel create animations of simple geometrical figures, which by their movement alone create the impression of intentional behavior. (Sketch adapted in style from original Heider and Simmel animations.)

As we have discussed in the MNS subsection, a lot of processing of visually perceived behavior is performed in area TPJ, in exchange with area PMC, in order to differentiate own from other's actions. The information about social intentions of an interaction partner, and the ability to continually differentiate this information, has been associated to be vital to the kind of mentalizing, which makes reasoning about latent mental states possible, i. e., mental states that need to be inferred, because they are hidden from our direct perception. Although the MNS goes a long way in processing information from a social scene, there is not much evidence for its involvement in long-term and more abstract mentalizing.

In a meta-analysis by Van Overwalle (2009), not only evidence for the involvement of the MNS was identified, but also the *MENT* – another functional network of brain areas in humans. It has been found to be strongly active during social interaction, when in order to make sense of someone's behavior, it becomes involved to understand their beliefs, personality traits, or behavioral intentions. With information from MNS-area STS, further processing is performed to identify action intentions that in cooperation with PMC can be differentiated from own actions in TPJ (remember fig. 2.1 with the MNS overview).

The area around *mPFC* is highly interconnected, even with regions not directly adjacent in the hypothetical processing hierarchy, e. g., there are connections to and from STS, TPJ, general PFC areas and even connections to thalamic and basal regions. mPFC has been associated with the long-term maintaining of response sequences in social behavior. It has been observed that *"neurons can continuously fire during an interval between an input and a delayed output. The mPFC stores these temporally disconnected events [. . .]"* (Van Overwalle, 2009, pp. 834).

mPFC: a highly connected brain area, also involved in MENT, associated with maintaining response sequences during social behavior, strechting over long time spans.

Fitting to the ability to represent long-term events and social behavior, different areas in the mPFC region have been associated with the ability to infer traits, or differentiate between close others or acquaintances. This suggests that the maintenance of different mental state representations, for intentions and beliefs, is a capacity where the mPFC is involved (see fig. 2.3 for a MENT overview).

In another review on the influences of different areas of the mentalizing network on social cognition (Schuwerk et al., 2014), it was found that posterior medial prefrontal cortex (pmPFC), when impaired, disturbs the ability to distinguish oneself from the other. Important for our later modeling approach is the ability to infer and track so-called social scripts, i. e., the sequence of social actions of all involved agents (including the self) that are adequate in the given social context.

Interestingly, another functional network of brain areas was discovered to be most active with the human participants at rest (*"lying quietly, with the eyes closed"* (Raichle et al., 2001, pp. 676)), without a task, called *default mode network* (Raichle et al., 2001; Raichle and Snyder, 2007) (which is also known as the *default system*). A meta-analysis

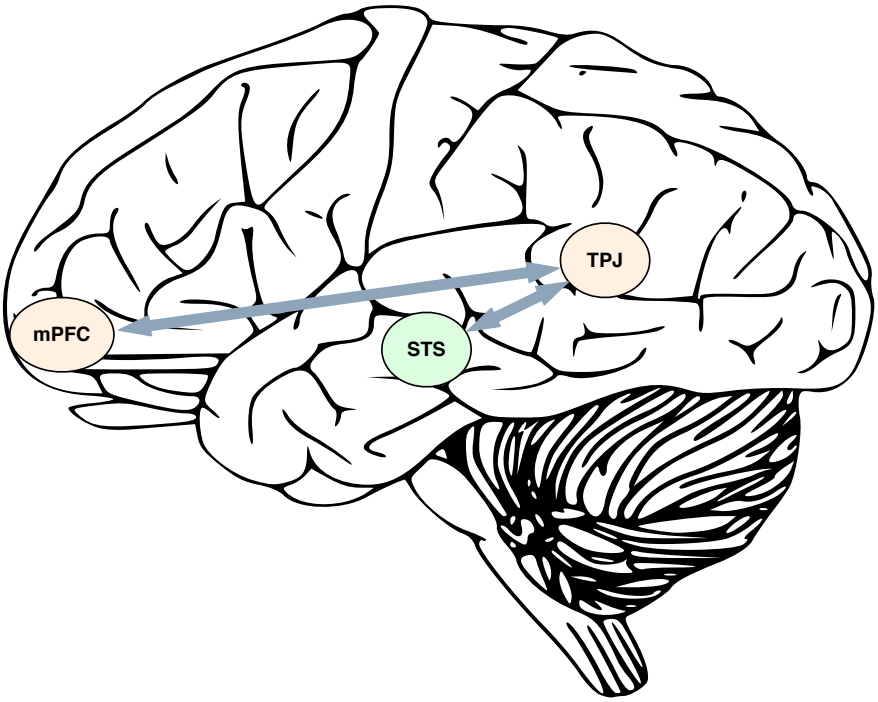


Figure 2.3: The mentalizing network consists of several functional brain areas that have been identified in a meta-analysis of over 200 fMRI studies (Van Overwalle, 2009). The ones most responsive during social cognition are schematically shown in this figure. Area TPJ is active during goal-directed action in social settings, also being involved in tasks demanding agency judgements. medial prefrontal cortex (mPFC) has been shown to be involved in inferring personality traits, theory of mind beliefs and inferring and tracking so-called social scripts of adequate behavior in the social context.

found these functional areas to be overlapping with areas typically involved in social cognition, which might imply that the default cognitive processing in humans is strongly predisposed toward social cognition (Schilbach et al., 2008).

2.2.3 *Interplay within the social brain*

Dark matter of social neuroscience: the missing mechanism underlying mental state attribution, from an interplay between mentalizing and mirroring processes, during reciprocal online social interaction.

The two functional networks of the social brain are not independent of one another and their interdependence and coordination will be the focus of this section. While a number of mechanisms have been hypothesized to underly social interaction, especially the mechanisms for reciprocal online social interaction, necessary for the human feat of mental state attribution, is still amiss and has been dubbed the *dark matter of social neuroscience* in a review by Przyrembel et al. (2012).

What is undisputed is that interacting with other agents assumed to be *intentional* is fundamentally different from interacting with non-intentional things, or objects (Gangopadhyay and Schilbach, 2012). Earlier analyses already argue that infants are aware of other's atten-

tion (Reddy, 2003). Further, this directedness of attention experienced by an infant merely from a *perception of attention* in others, can be seen as a precursor of a developing *self*. This would then develop first to a form of second-person perspective as a bridge between first-person and third-person perspectives. In this view, social interaction is ultimately rooted in the social cues and actions during real-time social interaction, e. g., social context shapes action kinematics (Becchio et al., 2010). This view is termed the second-person perspective, where the other's attending is *perceived* rather than *represented*, and the self as an object is *experienced*, rather than *conceived*. In a detailed discussion on the matter, Schilbach et al. (2013) argue that social and emotional engagement are necessary for a second-person approach to understanding other minds. As they have observed in *high-functioning autism (HFA)* patients, it is not the explicit mentalizing that is impaired in these patients, but rather the implicit process in direct social interaction, as it would normally allow them to automatically reorient themselves and integrate social cues.

This lack of automaticity during implicit mentalizing is described to be overwhelming for HFA patients, at times when directly engaged in interaction – in contrast to the patient being a passive observer (Schilbach et al., 2013). The second-person perspective calls for the necessity to investigate the procedural nature of real-time social interaction, which in neuroscience would mean the application of experimental setups that allow this, in order to investigate the pragmatic requirements to shed some light on the mechanistic underpinnings of dynamic social interaction, i. e., the dark matter of social neuroscience.

SOCIAL GAZE TRIGGERS SOCIAL COGNITION The human predisposition towards social cognition (cf. Schilbach et al., 2008) – the overlap of the default-mode network with MENT as discussed above – is also found in the mere judgement of whether somebody else's gaze is directed towards me. When uncertain, there seems to be a generalized tendency to judge another's gaze to be directed towards me (Mareschal et al., 2014).

Direct social gaze has repeatedly been found to trigger the functional stance necessary for social cognition that have been called an automatic form of implicit mentalizing, second-person perspective, or intentional stance. On the example of motor contagion (where perceived action primes own action planning), Becchio et al. (2007) describe the importance of gaze as a social cue. In their experiment they find that interference from motor contagion could only be elicited with the social cue, while in HFA children motor contagion could not be elicited under any condition.

Also, Myllyneva and Hietanen (2015) found a strong connection between typical responses to the mere belief of being attended to (skin conductance reponses, or a P3 response in Electroencephalography

(EEG)), and the participant's subjective view from a second-person perspective, i. e., as the object of another's attention.

For example, Ciaramidaro et al. (2014) recently found that social gaze leads to the attribution of communicative intent, which in turn differentially recruits the MNS and MENT networks, in processing the behavior of the interaction partner. Similarly, it was found that the mere presence of an other is able to activate MNS areas of the social brain (including inferior frontal gyrus (IFG) and PMC), while only direct social gaze with the interaction partner triggered the effective connection between IFG and mPFC (MENT areas) (Cavallo et al., 2015).

Although it seems that a key component to trigger this integrated form of social cognition is the bottom-up perception of social gaze, it was found that also a top-down regulation, from more explicit mentalizing areas of mere *beliefs about social attention*, can strongly modulate MNS areas, i. e., area STS was found to be influenced through the feedback coupling from areas mPFC and TPJ (Teufel et al., 2010).

So, what other factors, besides social gaze do trigger the activity of mentalizing areas during social interaction?

DOES VIOLATED ANTICIPATION INTERFACE MENT AND MNS? In the direct perception hypothesis (Gallagher, 2008), the perceptual accessibility to an interaction partner's intention is contrasted to the doctrine of surface behavior, which states that mental states are perceptually inaccessible. For example, Froese and Leavens (2014) found that precise imitation of other's actions is inhibited by correctly perceiving an interaction partner's intentions, unless the action is hard to interpret. They argue that the details are often overlooked, because they are not necessary for understanding, making the perceptual process automatic. This would make action imitation a learning response to unintelligible actions, involving more costly higher-level processing, and would also help to explain children's over-imitation when learning.

In a similar pointer to the on-demand involvement of higher-level processing, Bögels et al. (2015) found in a MEG experiment that during a referential naming task there was found no anticipatory mentalizing activity, but only on-demand mentalizing when an anticipation was violated and had to be accounted for. A plausible interpretation would be that generally expensive processing, like paying attention to the details of an interaction partner's action, or the search for a fitting explanation for her behavior, following a violation of my expectations, is only performed when the situation calls for it.

One such situation seems to be the direct interaction with somebody. Wykowska et al. (2014) argue that this allows for mentalizing activity to be triggered in order to form and retrieve beliefs about the interaction partner. In turn such beliefs can influence the processing of sensory information. This view entails a form of strategically taking sensory

information into account when necessary, in a form of social attention that to some degree controls the influence of sensory information.

2.2.4 *Self-other differentiation*

How does the human brain distinguish between our own and other's actions? Or to be more specific, how can we distinguish ourselves from others, so that we do not falsely attribute an action outcome to ourselves? Especially, during complex interaction scenarios with multiple interaction partners and overlapping behavior from oneself and others, such a differentiation can be difficult. Later, a possible solution to this problem will be discussed and become a core aspect of the social aspects of the model presented in this thesis. These questions are related to the general mechanisms that give rise to a sense of "feeling of control", agency, and "self". From here on: Sense of Agency (SoA).

SENSE OF AGENCY IN THE BRAIN A strong overlap of differential activity in the MENT and MNS networks can also be detected during SoA judgement tasks, which also underlies our ability to differentate our own from other's actions. Especially noteworthy: TPJ is a candidate to infer the agency of a social action, spanning areas STS which mainly responds to biological motion, to IPL which may respond to the intentions behind someone's actions. It connects to mPFC, which probably holds trait inferences, or maintains different representations of self-, or other-related intentions or beliefs. Generally, a person's SoA is believed to be influenced through predictive and postdictive processes (see next chapter in sec. 3.4) which, when disturbed, can lead to misattributions of actions, as in disorders such as schizophrenia (van der Weiden et al., 2015). Schizophrenia as a deficit of sensory attenuation, points to disfunctional precision encodings as a core pathology, i. e., the *confidence* of beliefs about the world (Adams et al., 2013). Precision is believed to be encoded in dopaminergic neuromodulation, and can as such be linked to the sensory attenuation effects during the attribution of agency in healthy subjects (Brown et al., 2013).

2.3 SUMMARY

Given the complex interactions that make up the dynamic involvement of mirroring and mentalizing activities within the social brain, it is no wonder that the underlying mechanisms have been called the dark matter of social neuroscience. To summarize, and with all that we discussed, let us return to the questions I formulated in the beginning of the chapter.

We have seen how we are able to get a very direct – and possibly predisposed – grasp on when we are part of a social interaction, and

what this entails in our means of processing prior information, and the context of the interaction also being involved.

Getting a grasp on the mental states of our interaction partner is the goal of the process of creating a shared understanding, while involving established common ground. We have reviewed the literature on how this process involves an implicit, and sometimes automatic mentalizing that reciprocally updates common ground, while also contextual information and prior information can influence our behavioral understanding top-down.

This process of reciprocal belief coordination seems imperfect, as it has been repeatedly shown that intial interpretations can linger, supported by contextual information, only being good-enough for the interaction to go on. But in general, misunderstandings and communication errors are not fatal for social interactions, but can be repaired. In fact the incremental nature of belief coordination is supported by findings that the higher-level process of mentalizing is triggered on demand, given violations of anticipation.

Of course, we still do not have a full understanding of all details of how social cognition is implemented in the brain, but we have now visited enough of the literature to be able to form a working hypothesis for the computational, and to a degree also of the algorithmic levels of analysis, along with multiple assumptions that we will visit in the following chapter.

3

MODELING FOUNDATIONS

We have visited the theoretical background surrounding the concept of belief coordination in human communication as well as the literature on the dynamics regarding repairs during such reciprocal interaction. Also, we looked at the social neuroscience literature, for the coordination of beliefs between interaction partners, and the necessary mechanistic processes of mirroring and mentalizing that make up social cognition. The present research focuses on finding mechanisms underlying the *intra-personal* dynamics in the social brain (MNS and MENT) and the *inter-personal* dynamics between interaction partners. This is done to the end of creating a dynamic interaction between computational models that can recreate the described behavior humans employ during belief coordination, including attempts for repair.

HPBU: the combined computational modeling approach, covering sensorimotor as well as mentalizing processes.

The combined modeling approach is called *Hierarchical Predictive Belief Update (HPBU)*, and covers two parts, related to the functional networks of the social brain: the sensorimotor part, and the mentalizing part.

We already covered many necessary elements of the model. First of all, modeling the sensorimotor part will allow for 1) dynamic and online perception and production of behavior. The mentalizing part will enable 2) prior beliefs, biases or social norms to influence future behavior, and 3) have beliefs about representations of behavior enable reasoning about other agent's mental states. These will need to be combined in 4) processes of perception and action to strategically guide the dynamic coordination of beliefs in social interaction.

In this chapter, we will first cover the necessary modeling theory for predictive processing (sec. 3.1.1). The next section covers the background necessary to understand mentalizing, or theory of mind (sec. 3.2). Integral to mentalizing is the self-other distinction, which we will discuss in the context of sensorimotor agency (sec. 3.4). Lastly, we will cover the related work on computational modeling of sensorimotor processes and mentalizing (sec. 3.5) and cover the contribution that the model presented here will make to the modeling landscape (sec 3.6).

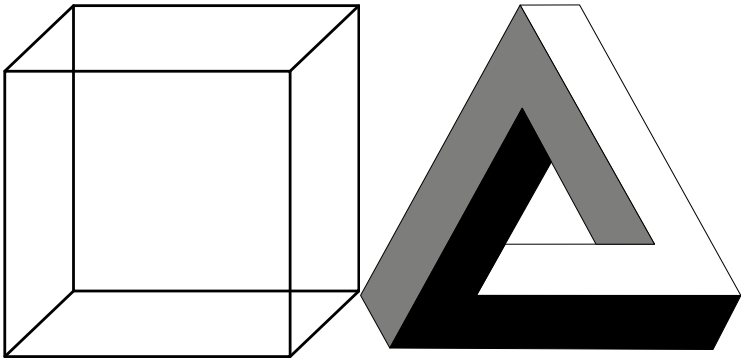
UNCERTAINTY The world in which we live in is one of sensory uncertainty, although our introspective perspective makes us believe that our perception is stable. Many factors can limit the reliability of our sensory information. Sometimes this is due to the sensorium at our disposal, e. g., the shape and position of our ears that make it quite impossible to detect the elevation of an auditory source. And

sometimes our preconceptions about the world bias our perception towards what we expect.

Humans and other animals need to minimize this uncertainty, in order to be able to perform the effective decision making. To be effective at this, new information needs to be combined with the experience we have gathered in the past. Getting this combination right is critical for our survival, e. g., as it decides whether we stay or run when we face a cat-like form while walking through a jungle. Thus it can be described as a core objective of our nervous system (Cox, 1946).

Let us have a look at how different approaches have so far tried to model the processes in the brain that allow it to perceive the world or act in the world under uncertainty.

In his monumental work, Helmholtz views human perception as statistical inference that takes sensory input and (unconsciously) infers probable causes (Helmholtz, 1867). The inferential process has become an integral part of cognitive psychology in a concept known as *analysis-by-synthesis*. That is, recognition is described as a process that compares priorly formed hypotheses with the sensory input with the goal of finding a matching hypothesis. This concept was taken up by Gregory



(a) Necker cube (b) Impossible triangle

Figure 3.1: (a) The Necker cube is a bi-stable illusion, because there are no cues as for its orientation. (b) The Penrose impossible triangle is an optical illusion that can only exist as a perspective drawing, but not in reality.

(1980), who analysed the comparability of percepts with scientific hypotheses. Therein the *signal* (information in the environment) was separated from the *data* (input data as processed by our sensorium), and the *hypotheses* (as inferred probable causes). In his analysis he does not shy away from the fact that hypotheses can be false, e. g., in cases of bi-stable percepts as in the Necker cube (Necker, 1832) (see fig. 3.1 (a) *). Also, distortions in the signal can be problematic, e. g., such as

* Necker cube by BenFrantzDale - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/wiki/File:Necker_cube.svg

a dim environment or auditory noise. Another interesting effect is discussed when it comes to distortions in the data that can be due to prior adaptations of the sensorium. An example would be different perceptions of heat in the left and right hand, when only one hand previously was calibrated to cold temperatures. Hypotheses can also be false when they are formed from inappropriate assumptions, as in the case of the Penrose impossible triangle (Penrose and Penrose, 1958) (see fig. 3.1 (b)†). This optical illusion can only exist as a perspective drawing, but not in reality. In recent years, the so-called *Bayesian brain hypothesis* suggests that a way in which these processes of the human brain can be modeled, is through probability theory (Knill and Pouget, 2004). It proposes that the human brain represents sensory information as a conditional probability density function, with every perceivable sensory information $\mathcal{Z}$ represented as a probability, given the evidence $\mathcal{I}$, as $P(\mathcal{Z}|\mathcal{I})$.

Bayesian brain hypothesis: a model of how the brain handles uncertainty, by means of conditional probabilities.

FOUNDATIONS OF GENERATIVE MODELS Such analyses led to the development of the *Helmholtz Machine* which describes its self-supervised learning as an approach that faces a statistical problem. It has the aim to maximize its likelihood by discovering a generative model that captures the given input's structure accurately (Dayan et al., 1995). A generative model learns how to produce the input that it gets presented with. The Helmholtz Machine makes a differentiation between perception and production, by learning a generative model for production, and a separate recognition model for perception. To train the model, the difference between an expected and a recognized probability distribution needs to be minimized. It is calculated using the so-called Kullback-Leibler Divergence (Kullback, 1959), which we will again meet later on. In a similar way (but without the statistical account), forward models have been considered to describe this form of generative models that allow for the generation of output that might exist in the brain. While for the recognition model, inverse models, or inversions of the forward models were proposed (Jordan and Rumelhart, 1992; Kawato et al., 1993). Combined, this allows for the generative model (or forward model) to influence the recognition model through prior activity (e. g., in a model of the visual cortex Kawato et al., 1993).

Early on, learning through feedback became a matter of investigation. Kawato and Gomi (1992) proposed on the example of four cerebellar regions of the brain, that learning might be due to error-feedback. Forward models that incorporate feedback in the form of a difference between predicted and perceived behavior. They are believed to be examples of generative models, which the brain may be using for both perception and action (Wolpert et al., 1995).

† Penrose triangle by Tobias R. – Metoc - Own work, Public Domain, https://en.wikipedia.org/wiki/File:Penrose-dreieck.svg

ON PREDICTABILITY The output generated by generative models can be utilized as predictions. Such predictions allow them to exhibit the same effect of *sensory attennuation* that has been observed in humans. Best described, it is the mechanism that inhibits the ability to tickle yourself. This is of importance, because when the difference between your predicted and perceived effect is minimal, so is the need for further processing of that sensory percept, which saves energy. Given the high energy consumption of our brain, compared to other organs of our body (Raichle and Gusnard, 2002; Sokoloff et al., 1955), it is vital for us to conserve as much energy as possible. Sensory attennuation is a strategy, based on predictability, which allows for that. For computational models of such processes, attennuation effects also have the effect of minimizing the need for computational ressources. This becomes necessary for systems that you build to run online, in real time if possible, so that you can have interactions between humans and machines.

Sensory attennuation: neuronal activation is inhibited if its reponse was predicted.

Predictability also has other advantages: taken to the level of interaction, it was observed that during duet playing on a piano, small delay variations strongly impact your ability to synchronize with your duet partner. The predictability of your own body makes yourself your best duet partner, because the prediction error from the delay variations are minimal (Keller et al., 2007). Similarly in a study where participants had to throw darts at a target board, they and others were recorded doing so. Later, a number of such videos were shown to them, and they had the task to predict where on the target board the dart would hit. Results show that own dart throws would be predicted with the highest accuracy (Knoblich et al., 2016). In another example for where the predictability of your body may give an advantage, is motor experience of specific tasks in sports. As it was shown in a study where participants had to judge the probable success of basketball throws from videos of partial throws (Aglioti et al., 2008). Expert basketball players were able to predict successful throws with much higher accuracy than novice players. These *"results suggest that only motor expertise endows the motor system with the ability to discriminate between erroneous and correct performance."* (Aglioti et al., 2008, pp. 1115).

3.1 PREDICTIVE PROCESSING AND ACTIVE INFERENCE

The modeling approach presented in this work has its foundation in these considerations about generative models, and the minimization of energy consumption. These rely on finding a way to estimate the worth and necessity of *computational work* being done, in order to minimize the energy consumption in the future. We will now explore a viable approach: the free-energy principle.

3.1.1 *The free-energy principle*

"Thermodynamic free energy" stems from statistical physics, where it describes the total energy available in a system to do work. The concept was introduced to the machine-learning literature by Hinton and Zemel (1994), who used the Helmholtz free energy to minimize an objective function to train an autoencoder.

Prediction error can be described as the difference between the way the environment actually is, and how it is represented in a system. Such a formulation can be seen in an information-theoretic light, where systems exchange information with the environment. As the environment can be increadibly complex, a system that lives inside it is not able experience it in a way that perfectly matches the system's representations.

Free-energy principle: living organisms, in constant exchange with the environment, need to resist the increase of entropy by keeping free energy minimial.

Friston (2013) describes the *free-energy principle* to not only encompass the information-processing of the brain. It is fundamental to whole systems, such as living organisms in exchange with their environment. Over time, living organisms have to resist the second law of thermodynamics, i. e., resist the increase of entropy. More specifically, living organisms such as us, maintain a model that represents the exchange with their environment. This model is tuned to minimize prediction error, or its information-theoretic homologue: *free energy*. Since this exchange with the environment is an ongoing process, the goal of a living organism must be to strive for an overall good model that minimizes free energy in the long term. That is, the free-energy principle requires the internal model of an organism and its actions to suppress prediction error and keep free energy minimal.

Free energy: the information-theoretic homologue of prediction error, i. e., the difference between a system's predictions and the actual state of the environment.

In the following, we will look at the free-energy principle at work in computational models, such as hierarchical predictive coding (Rao and Ballard, 1999). There, the minimization of free energy guides model selection – to predict probable hidden causes for the input – at higher levels of a hierarchy. Such a system should then be better at predicting future input at lower levels of a hierarchy, and thus suppress prediction errors. But the minimization of free energy does not only apply to model selection during perception, it also accounts for an organism's overt behavior. In the context of action, the minimization of free energy is not primarily done by model selection to infer the hidden causes of sensory states. Rather, the hidden causes are themselves influenced by action through *active inference* in order to make the hidden cause meet the prediction (Friston et al., 2010).

Active inference: a system's strategy to minimize free energy by actively changing the environment to meet its predictions.

Making correct predictions about hidden causes under uncertainty is tricky, so that the influence of prediction error on prior predictions has to be balanced carefully. This act of balancing needs to depend on the uncertainty itself, and is here described as the *precision weighting* of the prior predictions. This is like asking your own generative model: *"How well have my predictions performed so far?"*. Balancing the influence

Precision weighting: balancing the influence of prediction error on prior beliefs.

of prediction error on prior predictions by means of precision weighting can greatly influence the success of free energy minimization.

We will now turn towards more computational approaches that make use of the free-energy principle, such as hierarchical predictive coding.

HIERARCHICAL PREDICTIVE CODING Predictive coding was originally developed as a compression strategy. For example, in image data the compression of pixel color codes would depend on the predictability of neighbouring codes. When two neighbouring pixels would have the same color code, this is quite predictable and hence not newsworthy. When two neighbouring pixels were to be different, then this is not predicted by the previous pixel and hence is newsworthy. This is a very simple example, but might highlight the importance of predictability again, because in predictive coding only the newsworthy information is important and is further processed.

There is some evidence that the basic idea of predictive coding might be on the right track: in earlier work something similar was proposed as a kind of anti-hebbian learning in the form of *novelty filters*, where correlated activity leads to inhibition, rather than activation (Kohonen, 1983). Rao and Ballard (1999) propose a well-fitting model that accounts for the processing behavior of the visual cortex and they called it *Hierarchical predictive coding*. It accounts for multiple levels of processing where *"[e]ach level in the hierarchical model network (except the lowest level, which represents the image) attempts to predict the responses at the next lower level via feedback connections (Fig. 1a). The error between this prediction and the actual response is then sent back to the higher level via feedforward connections. This error signal is used to correct the estimate of the input signal at each level"* (Rao and Ballard, 1999, pp. 80). Their simulation results suggest that certain extra-classical receptive field effects can be interpreted as prediction error detecting signals. This could be an emergent property of the cortex, when using such a predictive strategy.

Hierarchical predictive coding: a hierarchical predictive model, where each level predicts its next lower level and only prediction error is communicated upward.

Similarly, retinal ganglion cells have been found to code differences to previous stimuli. Hosoya et al. (2005) propose that the behavior of those cells changes in order to adopt efficiently to new stimuli, in what they term "dynamic predictive coding". To that end the newsworthy information, as suggested by the original predictive coding account, is here transmitted onward to higher-level areas for further processing.

Also, in the architecture of cortical columns in the human brain, it was found that columns higher in the cortical hierarchy can inhibit activity at lower levels. Still, so-called *residuals* are the differences to the expected signal from higher-level predictions (Mumford, 1992). Given their findings they propose that *"an animal should not rest until it has 'explained' the full set of signals coming to it from the world, as far as*

its past experience allows, and must also be able to recognize when the signal indicates – because of variations beyond the normal limits – something never encountered before." (Mumford, 1992, pp. 246). This already highlights an important aspect of predictive coding as a generative model of brain function: the *explaining* of incoming signals, i. e., the newsworthy residual information needs to be explained and thereby *attennuated* higher up in the hierarchy. In predictive coding this newsworthy information is nothing else than prediction error and is attennuated – or explained away – at higher levels of the hierarchy. Also, the error is used to optimize future predictions. Friston and Kiebel (2009) combine the hierarchical predictive coding approach with free energy minimization as a means to describe the architecture of the brain's neocortex as a hierarchical generative model.

3.1.2 *Predictive processing*

Predictive processing, as defined by Clark (2016), combines a hierarchical system of a bidirectional probabilistic generative model, with the predictive coding strategy of efficient encoding and transmission. From this point on, I will thus talk about predictive processing and not just hierarchical predictive coding.

PRECISION-WEIGHTED UPDATES It is a tricky business to make predictions about hidden causes under uncertainty. Balancing the influence of prediction error on prior predictions right – by means of precision weighting – can greatly influence the success of free energy minimization. Precision weighting is described to be able to *"control the relative influence of prior expectations at different levels"* (Friston and Kiebel, 2009, pp. 299). Precision is also described as the gain on the prediction-error signal, which means that the higher the gain, the stronger the prediction error will influence prior predictions. Greater precision means that there is less uncertainty in the sensory data, leading to higher gain on the prediction error to be able to detect slight variations after the prior prediction was updated. Low precision biases the update process to preserve the prior prediction, while high precision lets the prediction-error signal drive future responses, by stronly influencing the update process.

In a similar notion, the adaptation to the variance of a given stimulus has been shown in the fly visual system, where neurons code for the variances of the luminance in the peripheral visual scene (Laughlin and Hardie, 1978). They describe this adaptation in terms of *efficient coding* (Barlow, 1961). That is, on the core assumption that neurons have a limited capacity to code for stimuli, the neuron's dynamic range is adjusted for the necessary stimuli to be most efficiently coded.

Another example often used, is that attention is modulated by precision weighting. Kok et al. (2012) manipulated independent prediction and spatial cues, in a Posner cueing paradigm setup (Posner, 1980), and found that attention reversed the inhibition effect of prediction upon the sensory signal. This means that congruent attention and prediction leads to an enhanced attention, while an incongruent simulus (that did not occur at the attended location) reduced the reponse measured in the primary visual cortex. Kok and colleagues argue that the effect of top-down prediction depends on attention. Taking attention into account, the sometimes different empirical observations – sometimes inhibiting and sometimes boosting the sensory data – can be explained. Press et al. (2019) discuss this supposed incompatibility in the general case (not only in visual attention). They propose a precision-weighting account that depends on the surprise of the sensory data, to account for the effect of empirical evidence, for both: inhibition and boosting of top-down predictions. Also on a very similar notion, Tatler et al. (2011) suggest a model of low-level salience based on uncertainty. They propose that action, perception, and attention should be seen in a more integrated manner, with attention and top-down driven influences from priors, learned from the environment.

These aspects are relevant, because this weighting of information is a key not only in the abstract sense of weighing prior information with new information. It should also be taken into account when we will later discuss the uptake of information from our social interaction partners, weighing it against what we presume about them.

ACTIVE INFERENCE Integral to predictive processing is also its approach to explaining, how action can occur, from a hierarchical system that maps from sensory effects to hidden causes. Such a system would, e. g., allow the imitation of human behavior, or enable the production of action in the first place.

This ability to translate visual information into action, or spawn action through thought, are both part of the scope of the so-called *ideomotor principle*. Voluntary actions describe this spawning of action through thought (going back to the analysis by James, 1890). The ideomotor principle describes the idea that if we think of an action's effects, like using a switch to turn on the light, this thought will resonate with representations in the motor system. The motor system then uses the motor program that has the strongest similarity with the intended action effect. Or in other words: *"when one wants to do something, the only requirement is to think of the intended action in terms of its ultimate distal result, and one need not care about the intermediate proximal steps required to realize it."* (Prinz, 1990, pp. 171).

Ideomotor principle: the motor system uses the motor program that has the strongest similarity with the seen or intended action effect.

Wohlschläger et al. (2003) describe that according to the ideomotor principle imitation does not depend on the observed movement as a whole. Rather, only different aspects of the movement are represented

and become active in a representational hierarchy, with the highest aspect becoming the main goal. Following the ideomotor principle this goal activates motor programs that most strongly correspond to that goal. Further they argue that meaningless, or partially seen movements can only be understood at lower levels of the hierarchy. The underlying mechanisms that implement the ideomotor principle have long been unknown. Recently, Pfister et al. (2014) identified the inferior parietal cortex and the parahippocampal gyrus as key regions for this type of anticipatory coordination of action.

In a similar way, active inference describes the process of spawning action. In active inference the environment is influenced through action to meet the given predictions that stem from held beliefs about the world (Adams et al., 2012). By making the environment meet one's held beliefs, they don't have to be updated by prediction errors from the environment. Following the assumption of predictive processing, the main flow of information is predictive, i. e., top-down. In active inference motor control can be seen as just top-down sensory prediction (Clark, 2016).

When active inference is seen as a process theory the distinction to previous models of motor control, is that in active inference we solely rely on each level's generative process to map from hidden causes to their sensory consequences. Without separate inverse models the generative process itself is inverted to predict the next steps at the next lower level, and thus, explains away or suppresses prediction errors. At the lowest level of the hierarchy, this process of suppression can take the form of triggering the production of actions, and change the environment. Thus in fullfilment of motor coordination – through a loop of sensorimotor feedback – such action can affect the inferred hidden causes, and minimize free energy (Friston et al., 2010).

3.2 MENTALIZING BACKGROUND

This section covers the theoretical background of the ability of mentalizing, i. e., creating a theory of mind to infer another person's mental states.

3.2.1 *Theory of Mind*

Since the time of Heider and Simmel (1944) many empirical investigations have improved our understanding of the mentalizing process, i. e., our ability to infer another person's mental states. The ability was termed ToM by Premack and Woodruff (1978), who investigated mentalizing in chimpanzee. Not long after, it became a focus of research in humans.

Dennett (1978) first proposed that a test for a false belief was necessary. Otherwise, it would be impossible to differentiate a person's

behavior to be in accord with reality or with her (potentially false) belief about reality. The *Sally-Anne false-belief test* (Wimmer and Perner, 1983) met this requirement. It was used to, e. g., investigate cases of childhood autism, which is a developmental disorder that impairs the understanding and coping with social environments (Baron-Cohen et al., 1985).

Stone et al. (1998) postulated ToM as a functional *system* that allows us to predict and interpret social information. They found that patients with damage to (bilateral orbito-) frontal cortex had difficulties in their ability to cope with tests that require more subtle social reasoning, such as the ability to detect a faux pas from a third-person perspective. That is, hearing someone say something to a third person, while not realizing that he or she should not have said it. The authors explain that this may be due to the need to represent two mental states: one for the person committing the faux pas, and one for the other person, who would feel insulted.

There is an ongoing debate about the nature of ToM, and its underlying mechanisms with two major contenders: simulation theory and theory theory.

THEORY THEORY AND SIMULATION THEORY Theory theory describes theory of mind to depend on a set of mental states, in the Belief, Desire, Intention (BDI) sense, and principles that guide their interaction. Thus, we formulate explanations and predictions about other's behavior using mental states, and generate behavior by combining the *theory* with prior information about the interaction partner (Gopnik and Wellman, 1992).

Simulation theory was developed to ascertain that the biology will allow the formation of beliefs, etc., which in themselves have causal properties (Gordon, 1986; Jeannerod, 2001). In that, simulation theory denies that we read other's minds by applying theory. Rather, it suggests that we use our own minds to understand others by putting ourselves in another's shoes, simulating their behavior. Also, beliefs are assumed to be (at least) similar in our interaction partner's minds, which allows us to think about other minds, using our own minds as a model to contrast to. This makes the simulation-theory account more akin to the mirror-neuron hypothesis (Gallese and Goldman, 1998), which has also been discussed to supposedly underly the whole of social cognition.

These accounts have long been viewed as mutually exclusive, but there exist integrated accounts of theory of mind, like direct social perception (Gallagher, 2008) which try to conciliate them, or are proposed as an alternative.

3.2.2 *The problem of recursion*

Already in the early days of research into ToM it became apparent, that mentalizing in complex social situations requires an ability to reason not only about the intentions and beliefs of your interaction partner. Also important is reasoning about the intentions and beliefs that your interaction partner might hold *about* your intentions and beliefs (with more rounds of recursive depth if necessary). Early studies showed that this ability develops during childhood, allowing for an increasing recursive depth during reasoning about intentions (Shultz and Cloghesy, 1981). Interestingly, this recursive depth in practice is not often applied to its full potential, as was shown by Keysar et al. (2003). They found that the ability to distinguish one's own from other's beliefs is not always applied when interpreting other's actions. Similarly, it has been shown by Devaine et al. (2014): participants in a game-theoretic task did test to employ a recursivity order of 2 most of the time, with 3rd order recursivity occuring negligibly. The authors even suggest an upper bound.

IMPLICIT AND EXPLICIT MENTALIZING Another set of strategies to handle reasoning about intention has been proposed and developed that makes use of a differentiation between an *implicit* and an *explicit* form of mentalizing (Frith, 2012; Frith and Frith, 2008). Where the implicit form underlies our ability to *automatically* perform in joint action tasks, and other non-verbal tests of non-verbal social cognition. It has been shown that this implicit mentalizing automatically influences gaze following, action imitation, and the tracking of other's knowledge (Kilner et al., 2003; Sebanz et al., 2006). Later, Frith (2012) proposes that during implicit mentalizing we employ a *we-mode* of mentalizing. Using the we-mode we are able to *automatically* infer and track the knowledge and intentions of others, without the need for a recursive reasoning about our own and our interaction partner's beliefs. In contrast, explicit mentalizing allows us to justify our behavior to our interaction partners, and meta-cognitively reflect on our own beliefs and intentions.

I will not further go into the developmental aspects of theory of mind. Rather, the scope of this work is to cover a *minimal* kind of mentalizing that is able to infer, and track the communicative intentions and goals during a non-verbal social interaction. Although, this may overlap with only part of the full abilities of adult humans.

3.2.3 *Conciliating theory theory and simulation theory*

We have to discuss where a predictive-processing perspective fits in the discussion about the nature of social cognition, i. e., can it shed light on the dark matter of social neuroscience?

Both perspectives (simulation theory and theory theory) are criticized for being *isolationist paradigms* (Becchio et al., 2010), in which interaction partners merely observe the other – only reacting to the mental states that they infer – instead of directly participating in the social interaction. This paradigm is challenged by a view in which social interaction is ultimately rooted in the social cues and actions during real-time social interaction, e. g., social context shapes action kinematics (ibid.). A more direct participation in the social interaction is put forward by the second-person perspective where the other's mind is *perceived*, rather than *represented* (Schilbach et al., 2013).

Also, Apperly (2008) argues that so far a discrimination between simulation-theory and theory-theory accounts was not found by neuroscience, as it was hoped for, and it probably never will. Further, the author finds no fault in the methods of investigation, but in the argument itself, which tries to differentiate between these accounts of theory of mind.

There have been attempts bridging the supposed gap, which rely on both: the folk-psychological inference, and the simulation – contrasting own from other's minds, in so-called *hybrid theories*. Gallagher (2015) compares such hybrid approaches – such as direct social perception – and argues for a pluralist approach that allows to make use of both, inferential and simulationist strategies.

de Bruin and Strijbos (2015) critisize that even proponents of direct social perception (Bohl and Gangopadhyay, 2013) still conform to the so-called *sandwich model* of social cognition (Hurley, 2008). The sandwich model of social cognition strongly distinguishes between processes for perception, cognition, and action. They discuss how direct social perception proponents disagree with the assumption that mental states need to be functional, instead focusing on their embodied grounding. They criticize that in direct social perception only the inferential process is cut out, which results in mental states. These discussions are contrasted with the *Bayesian Predictive Coding* approach that can do justice to both accounts. First of all, it gets rid of the distinction between perception and cognition by having perception to be unconscious inference (sec. 3). Also, it accepts that perception is driven by *theory*, i. e., priors in the brain. Finally, it undermines the distinction between perception and action by the close coupling of perception and action processes, as in *active inference*. Generally in the predictive coding account, the brain does not need to engage in deep inferential processing as long as the input is expected. Direct social perception also allows for inferential processing. But only when behavior is unexpected and cannot directly be responded to (Gallagher, 2008).

Given the hierarchical nature of the processes in the brain, for the following it is assumed that no *black box* of a cognitive process is at work. Rather, it is the predictive processing that conciliates hybrid

and sandwich model accounts by collapsing the distinction between perception, cognition and action. At work is only the *inferential resonance* of possible *perceptions* with the input, able to trigger prediction error correcting behavior, either to the end of correcting predictions, or to actively change the environment to conform to the expected perception.

3.3 MENTALIZING IN PREDICTIVE PROCESSING

3.3.1 *Event structures for mentalizing*

Now, we explore how social interaction, including mental states, can be represented in a predictive-processing hierarchy. Representations here need to account for the sequential nature of social interaction. Also, they need to comply to the free-energy principle, i. e., reduce uncertainty by attennuating prediction error.

Minsky (1974) developed frames and frame systems to describe a framework of memory. Frames are retrieved from memory to be adapted to fit reality, changing details as necessary. They are embedded in a frame system, a hierarchical structure where frames at higher levels represent things that are always true in a given situation and those at the lower levels are changing. This is one of the examples for segmenting knowledge and situations, while others are *schemata*, *scripts*, or *situation models* (which we will not all go into in the scope of this thesis). Rumelhart (1975) argues that stories consist of *episodes*, in which a protagonist has to achieve some goals, while each episode consists of a schema that contains steps towards that goal. The subgoals in the schemas which, able to contain schemas in themselves, make a recursive structure possible in order to flexibly achieve the goal.

It has long been unclear how such episodes can be learned, but a link to episodic memory is apparent, which is suggested to allow for *mental time travel* to remember past experiences (decoupled from reality) (Tulving, 1985). Zacks and Tversky (2001) argue that while the nature of our perception of the world is to some degree *continuous*, people need to understand it in terms of discrete events (also pointing toward the human bias to think in, and tell, stories). This discretization needs a way to segment the never-ending perceptual experience, and Zacks and colleagues argue that perception in itself determines how events are segmented. They develop the *Event Segmentation Theory* (Zacks et al., 2007), where events are segmented on boundaries that are detected from errors in predictability. New aspects that are unforseen in a given situation allow for segmentations, expanding the event structures, so that in the future similar situations are more predictable.

Such event structures – closely related to episodic memory – may at some level underly mentalizing. This is pointed to by imaging studies

that show reliable activations of the mentalizing region mPFC during retrieval from episodic memory (Hassabis and Maguire, 2007; Maguire and Mummery, 1999). Multiple meta-analyses on episodic memory and episodic simulation report an overlap with brain regions of the default mode network (Benoit and Schacter, 2015). They also find that there is a differential activation of brain regions, depending on the task at hand, requiring either more general reasoning or mentalizing. During mentalizing tasks, activation is strongly associated with the mPFC area (please refer to sec. 2.2.2) (Van Overwalle, 2011).

MODELING EVENT STRUCTURES The modeling approach presented in this work follows the evidence for the link between mentalizing and episodic memory to interpret social interactions in event structures, that will here be called *coordination sequences*. The proposed coordination sequences can be seen as schemas that contain segments consisting of mental state belief-attributions. The beliefs in the mental state attributions are updated from a sensorimotor part of the proposed Hierarchical Predictive Belief Update (HPBU) model (see next chapter for details). Coordination sequences make it possible to track the belief dynamics between interaction partners during belief coordination, over time, up until the interaction goal. The interaction goal is a final mental state that is to be reached.

Coordination sequences: represent event structures and make it possible to track beliefs over time.

The coordination sequence level is embedded in a hierarchy where, at an even higher level, coordination sequences can be collected into clusters of sequences with similar interaction goals. These clusters of coordination sequences are here called *goal states*, as they collect sequences of similar outcome. Having different sequences toward the same outcome available enables the goal-state level to predict the outcome of an interaction (in the form of an interaction goal).

Goal states: represent clusters of coordination sequences with the same interaction outcome.

Further below, both representation levels will be described in more detail, when we discuss the computational model of mentalizing. As the literature suggests, similar event structures could also exist for general reasoning purposes, outside of mentalizing, but this is beyond the scope of this work.

3.3.2 *Minimizing free energy of beliefs and intentions*

Coordination sequences have access to information from sensorimotor processing, along with information about the agency of an action that, as we will discuss soon, allows for the differentiation of own from other's behavior. Given the access to such information, beliefs can be attributed to oneself or an interaction partner, and under the segments contained in the coordination sequences, a fitting segment to the mental state can be identified, along with possible interaction goals, or a specific goal of the inferred coordination sequence.

This function of identifying possible interaction goals can be described as a mechanism for modeling the so-called *affordance competition*. Described briefly, it was introduced as a possible cortical mechanism for action selection (Cisek, 2007), and has since been developed to also cover more abstract – or higher-level – goals (including ways to create new affordances) (Pezzulo and Cisek, 2016). Thus, the here proposed mechanism is one of affordance competition, for the case of social interaction.

Affordance comptetition: a cortical mechanism for action selection, where possible goals afford interaction with the environment.

Given that coordination sequences are embedded in a predictive processing hierarchy, they must comply to the free-energy principle, minimizing prediction error, and hence free energy. The here presented idea for a form of social predictive processing was developed from the discussion in Friston and Frith (2015b). They interpret the term *hermeneutics* to apply to the problem of interpreting the intention from verbal and non-verbal communication. They model the problem in the form of two birds that can hear and produce a song. In that model scenario, they describe their solution to the problem of neural hermeneutics based on active inference, where action fulfills our predictions about our own behavior. This way one's own action can be attenuated, and the correctly predicted bird song, of the other bird as well. An important point made in this paper is that the simulated birds can hear themselves, not only their interaction partner. This underlines the necessity to be able to correctly predict and attenuate own actions, in order to tell them apart from that of an interaction partner.

In a companion paper, Friston and Frith (2015a) make their view of communication clearer: They point out that communication is based on a shared narrative, in which interaction partners reciprocally attend to and attenuate sensory input in a back and forth manner. They argue that this back and forth can lead to a synchrony between the brain states (in the form of probable hypotheses) of interaction partners. A central problem of communication is the infinite regress induced by modeling your interaction partner (while she is modeling you). The regress can be (partially) evaded if both possess a similar model, as this would minimize surprise, maximize predictability, and would allow for synchronized brain states. This leads both interaction partners to generate successful predictions about each other's mental states. In a way, the described synchrony even allows for the interacting brains to bypass the perceptual loop that spans the interaction partner. Thereby, they effectively predict themselves.

Thus, if predictions become shared between interaction partners in the form of overt behavior through active inference, uncertainty can be minimized. Still, the interaction described in both papers is schematically similar (but rather primitive in comparison) to human belief coordination (as described in ch. 2).

PRIOR EXPERIENCE WITH THE INTERACTION PARTNER Common ground (see sec. 2.1.1) has been described as the prior information that you bring into a social interaction. This can be information about the world, and this can also be prior experience you share with your interaction partner.

If there is such prior experience that you share, it should affect your inferential processes, and your predictions about the other's behaviors and beliefs. For example, if you know that you and your conversation partner previously discussed where best to get dinner, then in a later conversation you will probably know what restaurant is meant when she picks up the conversation, asking you: *"When do you want to meet at the restaurant?"*. On a more fundamental perspective, the common ground established between you and your interaction partner influenced your inferential processing to minimize uncertainty about which restaurant is meant.

In several imaging studies the dorsal mPFC area in the brain was associated with inferring personality traits and beliefs (Van Overwalle, 2011). As we also saw, the episodic memory literature also finds a strong overlap with area mPFC during episodic memory retrieval (Hassabis and Maguire, 2007). Thus, this is probably a good candidate to suggest an area from which other mentalizing and action perception processes could be influenced by prior knowledge from mental state beliefs and personality traits, i. e., a form of *person model* (PM).

Later, the computational person model, and how it influences inferential processing during social interaction, will be discussed further.

3.4 INFERRING THE SELF FROM SENSE OF AGENCY

How does the human brain distinguish between information about ourselves and others? As shortly introduced, the social brain is strongly involved during judgements of agency and the differentiation of own from other's actions (see sec. 2.2.4). Generally, a person's SoA is believed to be influenced through predictive and postdictive (inferential) processes which, when disturbed, can lead to misattributions of actions, as in disorders such as schizophrenia (van der Weiden et al., 2015). The aim of the following discussion of the predictive and postdictive processes, is to identify their underlying mechanisms. These will be modeled as a combined representation, which forms a cue for SoA of perceived behavioral outcomes.

3.4.1 *Predictive process in sense of agency*

The *predictive process* makes use of people's ability to anticipate the sensory consequences of their own actions. It allows to attenuate,

† This section on the mechanisms of sense of agency for self-other distinction was adapted from a section in a previous publication in Kahl and Kopp (2018).

i. e., decrease the intensity of predicted incoming signals. Using this mechanism, it enables people to distinguish between (predicted) self-caused actions and their outcomes, and those (unpredicted) actions and outcomes caused by others.

One account to model these processes is based on inverse and forward models (as discussed in par. 3), to account for disorders of awareness in the motor system and delusion of control (Frith et al., 2000). This account suggests that patients suffering from such disorders of awareness can no longer link their intentions to their actions. They still can become aware of the sensory consequences of an action, but may find it problematic to integrate them to the intention underlying the action. This would make it difficult to ascribe actions to oneself or another agent, making misattributions more likely.

Research on schizophrenia has shown that reliable and early self-other integration and distinction is important not only for the correct attribution of SoA, but also for the correct attribution of intentions and emotions to others in social interaction (van der Weiden et al., 2015). Weiss et al. (2011) also showed that there is a social aspect to predictive processes that influence SoA, by comparing perceived loudness of auditory action effects in an interactive action context. They found that attenuation occured also in the interactive context, comparable to the attentuation of self-generated sound in an individual context.

Another aspect of the processing of differences between predictions and feedback from reality is the intrinsic robustness and invariance to unimportant aspects in the sensory input. Our model concerns itself with allowing to act in (and perceive) the ever-varying nature of its environment, while being able to attenuate the prediction errors that aren't surprising enough to lead to any form of adaptation (see sec. 3.1.1). That this is also likely true for temporal prediction errors was found in Sherwell et al. (2016) who, using EEG, saw significant N1 component suppression in predicted stimulus onset timings. Consistent with this perspective is work by Rohde and Ernst (2016), who investigated if and in which cases we can compensate for sensorimotor delay, i. e., the time between an action and its sensory consequence. They find that if an error signal (a discrepancy between an expected and an actual sensory delay) occurs we recalibrate our expectations only if the error occurs systematically. This kind of temporal adaptation is a well studied finding (e. g., Haering and Kiesel (2015) for sense of agency or Cunningham et al. (2001) in motor control).

It is the unexpected, unsystematic and sudden temporal deviations in sensorimotor processing that we will focus on next.

3.4.2 *Postdictive process in sense of agency*

The *postdictive process* relies more on inferences drawn after the movement, in order to check whether the observed events are contingent and consistent with specific intentions (Wegner and Wheatley, 1999), influenced by higher-level causal beliefs. Such beliefs may stem from mental states of interaction partners, inferred through theory of mind. But also expectations from folk physics might be at play, which Daniel Dennett describes as: *"Folk physics is the system of savvy expectations we all have about how middle-sized physical objects in our world react to middle-sized events. If I tip over a glass of water on the dinner table, you leap out of your chair, expecting the water to spill over the side and soak through your clothes."* (Dennett, 1989, pp. 7-8).

Temporal aspects of action-outcome integration seem to be important for this inferential process. For example, it was shown that increasing action outcome delay decreases feeling of control (Sidarus et al., 2013). Colonius and Diederich (2004) describe a model for the improved response time in saccadic movements towards a target that is visually and auditorily aligned. Their *time-window-of-integration* model serves as a framework for the rules of multisensory integration. This integration occurs only if all multimodal neural excitations terminate within a given time interval. In van der Weiden et al. (2015) this time interval of integration is taken as a solution to a problem posed in the classic comparator model of motor prediction. The brain needs to integrate action production signals with their predicted outcomes, which can be perceived via multiple sensory channels. Such action-outcome integration needs to account for the different time scales in which outcomes of actions may occur.

A point not taken into account by Colonius and Diederich was how such an integrating mechanism knows how long it has to wait for all action outcomes to occur. Hillock-Dunn and Wallace (2012) investigated how these temporal windows for integration – which have been learned in childhood – develop through life. They analyzed responses to a judgement task of a visual and an auditory stimulus to occur simultaneously in participants, with ages ranging from 6 to 23 years, and found an age dependent decrease in temporal integration window sizes. A possible assumption would be that a wider window of integration can be associated with un-predictability and greater variance in action outcome timings. A decrease of the integration window size may be due to an adult person's experience advantage about effects their actions may have on their environment, or the mere improved predictability of their full grown bodies.

Such an integration of an intended action with its predicted consequences, learned through associations between action events, can lead to an interesting phenomenon, often reported in the SoA literature. In this phenomenon, integration can lead to the effect that predicted

action consequences can be perceived to occur at the same time. This phenomenon is called *temporal binding*, or also *intentional binding*, when the effects of an intended action are predicted and are perceived as occurring closer together as unintended actions (Haggard and Clark, 2003; Haggard et al., 2002). Temporal binding is one of the measures often used to test for study participants' SoA.

In sum, by and large, there are two processes that can inform SoA, and hence can help to distinguish actions of self and other, in social interaction. A predictive process is based on (assumed or given) causes of the action, e. g., a motor command is executed and a forward model is used to predict the to-be-observed sensory events.

A postdictive process works with features of an observed action outcome and applies higher-level causal beliefs and inferential mechanisms, e. g., a given intention to act, or temporal binding, to test the consistency of the action outcome and infer a likely explanation.

3.4.3 *Integrating sense of agency*

How are these two processes integrated to inform SoA, and what if their cues are unreliable? When disorders of SoA were first studied, the *comparator* model was the first proposal concerning its underlying mechanism. This was soon questioned as the comparator model failed to account for external agency attributions. It was argued that its evidence has to be weighted and integrated with more high-level sources of evidence for sense of agency (Synofzik et al., 2008). The weighting and integration of such evidence cues was studied by Moore et al. (2009), who found that external cues – like prior beliefs – become more influential if predictive cues are absent.

Neurological evidence for a differential processing of cues that inform SoA comes from Nahab et al. (2010). In an imaging study, they found a *leading* and a *lagging* network that both influence SoA, prior to, and after an action. The leading network would check whether a predicted action outcome is perceived, while the lagging network would process these cues further to form a SoA that is consciously experienced.

Further, an EEG study found evidence for separate processing areas in the brain (Dumas et al., 2012b). They triggered predictive and postdictive cues in two tasks. One induced an external attribution of agency, while the other used a spontaneuous attribution condition. It seems that in order to generate SoA, both systems do not necessarily have to work perfectly together. Instead, the SoA might be based on a weighted integration of predictive and postdictive cues, depending on their respective precision (Moore and Fletcher, 2012; Synofzik et al., 2013; Wolpe et al., 2014).

Furthermore, the fluency of action-selection processes may also influence self-other distinction, because the success of repeatedly predicting the next actions seem to accumulate over time to inform SoA (Chambon and Haggard, 2012; Chambon et al., 2014). This action selection *fluency-effect* seems to contribute prospectively to a sense of agency, similar to a priming effect.

We have now discussed the literature on how the information about our own body and actions can influence wheather we can differentate other's from our own self-produced actions. This information is used in a two-process approach, one predictive and one postdictive, i. e., one evaluates the action based on the prediction, while the latter evaluates it in a broader context. The evaluated sense of agency, resulting from these processes, is then integrated over time. This integration happens in the information-theoretic context (free energy and precision) of the current action-perception loop, and the broader temporal context that HPBU is in.

3.5 RELATED WORK IN COMPUTATIONAL MODELING

One aim of this thesis is to create a computational model of processes in the two parts of the social brain that lead to dynamic production and perception behavior, allowing the coordination of beliefs of multiple agents. As described at the beginning of this chapter (see ch. 3), necessary elements consist of:

1. allowing for dynamic and online perception and production of behavior
2. enabling prior beliefs, biases or social norms to influence future behavior
3. having beliefs about representations of behavior enable reasoning about other agent's mental states
4. applying processes of perception and action strategically to guide the dynamic coordination of such beliefs.

We now visit and review literature on computational cognitive modeling of the necessary processes.

3.5.1 *Kinds of models*

Bayesian (or probabilistic) modeling is one of the big four approaches to cognitive modeling, with the others being connectionism, rule-based approaches, and dynamic systems. All have their strengths and weaknesses, and have their problems to tackle, with one being neural plausibility, and the other – which is more specific to Bayesian

modeling – being rationality. All cognitive modeling eventually has the goal to model the brain, processes within, or the interaction thereof, in a plausible manner, i. e., create a reductionist account of cognition.

Bayesian modeling and rule-based approaches have the same problem: answering how rules and inference are implemented in the brain. Connectionist accounts are deemed more neurally plausible, as they are supposedly based on models of the same building blocks as the brain, i. e., artificial neurons. But they also lack answers, e. g., when it comes to how new neurons (due to neurogenesis) can partake in an artificial neural network, or how the fine detail of the biology of neurons influence their behavior.

Connectionist models can be described as a more bottom-up approach, while probabilistic models take the top-down perspective when it comes to modeling the neural mechanisms underlying mental processes. Griffiths et al. (2010) argue that the top-down perspective is more flexible when it comes to exploring the representations and biases underlying human cognition. Rather different is the perspective by McClelland et al. (2010), who argue that probabilistic models sometimes serve as misleading abstractions with no real basis in the actual process. They argue that connectionist and dynamic systems approaches are better suited to explain the actual mechanisms that give rise to cognition.

Commenting on these two arguments, Marcus argues that *"[g]enuinely adequate theories must borrow from each of these traditions – connectionism's emphasis on development, and on how complex cognition derives from the actions of relatively simple low-level units; Bayes' emphasis on reverse engineering (shared with evolutionary psychology). But both groups take a one-size-fits-all approach that isn't warranted by the data. Only by severing commitments to extreme empiricism and excess adaptationism can we hope to span the chasm between low-level neural circuits and higher-level cognition, in a fashion that is faithful to the reality of our evolved psychology, quirks and all."* (Marcus, 2010, pp. 2).

Another big problem for Bayesian models of human cognition are limits of rationality. On the one hand, Bayes' rule is inherently rational, while on the other hand, humans have repeatedly been shown to not be, when it comes to decision making under uncertainty (cf. Shafir and Tversky, 1992). How prior and likelihood information is to be optimally combined with respect to ther uncertainties is described in Bayesian statistics, so a posterior probability estimate can be formed. A way of representing uncertainties in the brain could be in the form of distinct populations of neurons, as suggested in the form of probabilistic population codes or relative timing effects (e. g., Knill and Pouget, 2004 and for a review see Vilares and Kording, 2011). If it were that uncertainty is taken into account at every state of the cortical processing hierarchy in the brain, then we would be able to

talk about a truly Bayesian brain. But given the sometimes irrational behavior of humans, Clark (2016) describes the human brain to be not *Bayes optimal* but rather, optimal at taking uncertainties in the information from our own senses into account, making humans act as *rational Bayesian estimators* (e. g., Berniker and Kording, 2008). In that sense, the predictive processing account claims that the brain *approximates* Bayesian inference, while uncovering the hidden causes of the available information.

Let us see how these different modeling approaches have been applied to handle the uncertainty and dynamics of motor coordination, behavior recognition, or coordination of beliefs that form a theory of mind.

3.5.2 *Models of motor coordination*

Motor control has to solve two problems to allow for comprehensive interaction with the environment. One problem is to learn action sequences that allow for understanding the goal of an action, by mapping from a perceived action (in an extrinsic coordinate frame), onto a description of muscle movements (in an intrinsic frame). The second problem is to reach an action goal by activating the appropriate muscles in sequence, to produce the necessary movement. This is a hard problem, because most often many competing solutions exist, for the mapping from an intrinsic to an extrinsic coordinate frame.

In the early days, often the planning of an optimal trajectory towards a goal was assumed, applying models that tried to find an optimal sequence of muscle activations (Kawato, 1999). The calculations to find optimal action sequences was done offline, before the action event started.

The high variability that was observed in natural movements, even under conditions of repeated tasks, led to another strategy (Todorov and Jordan, 2002). One which allowed for variability in redundant task dimensions during action production, where feedback showed that it doesn't interfere with reaching the goal. A strategy related to the dynamical systems view of motor coordination (Kelso, 1995). They coupled an optimal feedback controller with the controlled motor plant (the coupled system of moving joints) to produce a dynamical systems model that allowed to coordinate movements to reach specified targets. Thus, *motor coordination* describes the flexible and feedback-involved control of movements of a motor plant to the end of reaching a movement goal.

Motor coordination: the flexible and feedback-involved control of movements to the end of reaching a movement goal.

Paired forward- and inverse models have been proposed to model general purpose motor behavior. Also they should be able to learn to

† The matter of how active inference could implement a comprehensive motor coordination loop was also discussed in a previous publication in Kahl and Kopp (2018).

infer motor behavior, while a switch to an appropriate inverse model can be performed during perception. These models incorporate ways to handle probabilistic uncertainty during motor execution. This is necessary to account for the sometimes considerable time delays in the central nervous system for signal transduction and the motor execution itself. It is also used to identify the "responsibility" (likelihood) of a forward model to accurately model the perceived behavior. Extended into a hierarchy of pairs of forward and inverse models, Wolpert et al. (2003) call their model MOSAIC, and discuss and explore the similarities between motor control and social interaction. They stress that in principle MOSAIC should be able to account for motor control and social interaction, although this has not been proven in practice. MOSAIC makes strong use of sensory feedback, comparing it to a predicted action. Feedback would influence the embedded forward and inverse models, in order to optimize future action. Here, feedback describes the proprioceptive feedback coming from so-called spindle-cells surrounding the muscles as well as the joint positions that are perceived visually.

More generally, in comparison to the fordward models used in models of optimal control, as Friston (2011) highlights, the generative models in perceptual inference are different and should not be conflated. He argues to replace the optimal control problem with active inference, thereby making it an inference problem over motor reflex arcs. As we will see later, this approach can use the information from the extrinsic coordinate frame to circumvent the need for detailed planning of muscle activations.

The recent revival of connectionist approaches was grounded in increased processing power of modern computers, parallelization, and also the success of new Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) architectures. Especially, so-called Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) found successful applications, e. g., also in sensorimotor control. For example, a RNN was applied on a complex robot platform by Schilling and Cruse (2012), to perform hexapod walking. This RNN was trained to create an internal model about the robot body, reacting to the world. Here, the internal model could also be used as a forward and inverse model. It not only allows to infer the forces applied to the robot, but it also allows to optimize the robot behavior by making predictions over future actions.

Also, another promising example should be considered, as it rests on similar assumptions as the theoretical basis for the modeling, as presented in this thesis. Butz (2016) proposes that LSTM networks could be applied as predictive, generative models to be applied in a top-down manner, generating sequences, such as those underlying motor control. This approach was further developed into the REPRISE model, which is described as a retrospective and prospective inference

scheme. It takes into account the recent sensorimotor context retrospectively, selecting best-fitting event-models, to then prospectively predict the next motor activations (Otte et al., 2017). REPRISE was successfully trained to infer and predict different sensorimotor contingencies in a partially observable scenario, where different kinds of space ships have to reach a target (Butz et al., 2018). The model plans ahead multiple time-steps of sensorimotor control, in order to reach its goal, successfully taking into account the sensorimotor contingencies presented by the scenario and the space ship.

Of course, these models of motor coordination are missing a mentalizing perspective that might allow to infer mental states which might explain the perceived behavior differently. Also, the adaptation to new movements and situations might only be limited, as deep connectionist models with many layers are known to be data hungry and might not generalize to other types of movement. Nevertheless, the sensorimotor part of HPBU employs ideas from the motor coordination literature. Now, we turn to computational models of social interaction, as these play a role in the mentalizing part of HPBU.

3.5.3 *Models of theory of mind*

We find two kinds of models in the literature on computational models for social interaction. One kind of computational approaches understands social interaction through observation, i. e., inferring beliefs about other agents through observing their behavior. Also, there are a few computational models that describe the direct social interaction through modeling agents that actively participate in an interactive task. These allow to investigate the underlying dynamics necessary for joint action and belief coordination.

Here are some examples for inferring beliefs from observing behavior: The approach of inferring beliefs of other agents has its clasical origins in symbolic reasoning approaches, most prominently the BDI model of computational agents (Rao and Georgeff, 1995). With its reliance on symbolic reasoning and hard-coded rules, models based on this framework often provide well defined representations of social situations, but which are handcrafted very specifically. The symbolic and rule-based approach leaves virtually no room for variability as would be needed to interpret real-world behavior.

To better cope with uncertainty, probabilistic approaches have been proposed for recognizing an agent's behavior from observation (also referred to as *plan recognition*) early on. This has traditionally been done, either by directly translating and representing the plan recognition problem into a Bayesian network, and to represent possible explanations in a probability distribution (Charniak and Goldman, 1993). Also, this can be done by applying a general Dynamic Bayesian

Network (DBN) to allow for temporal reasoning (Dean and Kanazawa, 1989) and, e. g., predict a player's next action in an adventure game (Albrecht et al., 1998). For example, Han and Veloso (2000) applied multiple networks of DBN's specialized cousins – Hidden Markov Models (HMMs) – to recognize robot behavior, where each HMM infers a single plan behavior with their hidden variable.

Taking a step beyond plan recognition, Pynadath and Marsella (2005) describe a method called PsychSim. It is used to simulate how agents would take other agents into account. It would not only represent beliefs about the world, but also assume them in other agents and include these during decision making, though this model does not infer mental states of agents from their behavior.

Again, with a focus on inferring the beliefs of agents, Baker et al. (2011) propose a framework called Bayesian Theory of Mind (BToM) which is based on a process called Bayesian inverse planning. It assumes that human reasoning on observed behavior is based on generative processes that allow them to rationally generate behaviors, given the state of the environment, their beliefs, the beliefs of other agents, and so on. Further, they assume that humans infer the goals of others by inverting this generative process.

Following the idea of inverse planning, Pöppel and Kopp (2019) describe an approach following the bounded rationality idea, i. e., that human decision making is only optimal given the information we have. Further, they follow the assumption that humans strive for being good *satisficers*, i. e., to *"choose options that are good enough to satisfy a given need instead of actually evaluating all possible options in order to choose the objectively best one"* (Pöppel and Kopp, 2019, pp. 2). In the context of inferring agent's goals from observing their behavior, they present a *switching model* that provides specialized models of different complexity that can be applied when simpler models can no longer provide meaningful explanations. They found that specialized models can outperform complex models, where applicable, and found the switching model approach to outperform the naive (or more general) approach.

In light of recent innovations in the connectionist approach, called *Deep Learning*, progress was made to infer a theory of mind from observing another agent's behavior. Rabinowitz et al. (2018) trained recurrent neural networks and extracted high-dimensional embeddings onto a two-dimensional plane. This way, clusters could be found that could be interpreted as an agent's observable behavior, mapped onto a representational space of mental states. It requires large amounts of synthetic training data, which makes it unpractical to use for human behavior, but the network could predict an agent's behavior from these inferred mental states.

These symbolic and rule-based approaches are limited in their ability to infer mental states from behavior that is grounded in more dynamic

and often quite variable real-world movements that make it necessary to integrate behavior information over time. Also, the coordination of beliefs, while taking into account the mental states of multiple agents, are problems that have yet to be solved.

3.5.4 *Models of direct social interaction*

In contrast to the approaches above, models of direct social interaction try to understand the dynamics underlying social interaction between agents that stand in direct contact. Direct contact means they are either directly coupled physically, or in a more loose manner, coupled through communication, in a joint action task.

Only very few computational models of direct social interaction can be found in the literature. One model of joint action was developed by Pezzulo and Dindo (2011), which was used to show how shared representations can help solve interaction problems. The model was based on a probabilistic account with two levels of Dynamic Bayesian Networks, where at the lower level the same "circuitry" is used for action production and perception, in the form of forward and inverse models of activation. The higher level model provides a "motor simulation" process, which is guided by prior intentions, in order to allow for action intentions to be inferred. The interplay between both levels is in both directions, allowing prior intentions to bias action perception, and recognized motor primitives can act as abstract representations of the joint task with an interaction partner.

Recently, similar computational models were used to investigate the dynamics of turn-taking in conversation, with multiple agents interacting. They showed that sensorimotor communication and prediction of the intention behind the other's action brings the best results, minimizing the gap between turns (Donnarumma et al., 2017). Similarly, Sadeghipour and Kopp (2010) present the Empirical Bayesian Belief Update (EBBU) model, a probabilistic model that implements a mirroring-based account of the perception and production of iconic gestures. In this model, a hierarchically organized representation of motor knowledge is used during action perception by forward models that formulate probabilistic expectations about possible continuations of the observed gesture. The same representation is used for action generation, with probabilistic interactions between both processes to model, e. g., priming and resonance effects, and it is expanded by way of inverse models when an unknown action is encountered. They describe how this works during interaction with a human interaction partner, who interacts with a virtual agent that is equipped with the EBBU model.

Also, in a first prototypical model of the work presented in this thesis, a version of the EBBU model (ibid.) was extended to communicate with a heuristic mentalizing model (Kahl and Kopp, 2015). In

that paper the question was investigated how mere action observation needs to be complemented by higher order mentalizing, and how those systems need to interact, in order to account for the dynamic inter-agent coordination mechanisms that are required for successful communication. Using an interactive communication game with two virtual agents (both equipped with the mirroring and mentalizing model) it was investigated whether 1st order mental state attributions are sufficient to infer the information, necessary to successfully act towards a communicative goal, or whether higher-order theory of mind can give a distinct advantage. These results demonstrated that mentalizing affords interactive grounding, and thus makes communication more robust and efficient.

What is missing in these previous accounts are approaches to modeling the simultaneous production and perception of behavior. This is necessary not only to confirm the correct production of own actions, but also to be able to observe behavior of other agents, even during own action production. Also, the extended EBBU model implemented mentalizing processes as simple heuristics, without being put on the same foundation of handling uncertainty, as the sensorimotor processes.

Also, there are some proposed computational models on direct social interaction without implementation and evaluation. For example, Brandi et al. (2019) propose a computation model based on predictive processing to be used to analyze data from ecologically valid and interactive study designs. They also review the literature on how virtual reality can help to simulate social interaction to better study so-called *social agency*. Social agency can be defined as a form of feeling of control within social situations, e. g., being able to predict an interaction partner's contributions to a conversation. Similarly, Fotopoulou and Tsakiris (2017) discuss the connection between *interoception* and social interaction. Interoception describes a sense of the internal state of one's body (Khalsa and Lapidus, 2016). They argue that an awareness of self (through interoception) is crucially linked to the experience of dynamic social interaction. Rightfully, they make a call for more work on computational models of social interaction, in order to reach mechanistic explanations for what makes it so special.

Social agency: a feeling of control within social situations.
Interoception: a sense of the internal state of the body.

3.5.5 *Models of interactive brain dynamics*

When it comes to synchrony between interaction partners, there is not only evidence for synchronizing behavior. Also, the synchrony between activity in the brains has been suggested, between interaction partners who's behavior aligns. Such a correlation between behavioral and brain synchrony has been found by Dumas et al. (2010) who made hyperscanning EEG recordings.

Later, Dumas et al. (2012a) created computational models of weakly coupled non-linear oscillators – the so-called Kuramoto model (Kuramoto, 1975). A model commonly used in the study of synchronization phenomenon in physics. The oscillators were coupled to conform to the points of localized activity from the EEG recordings, to represent one brain. Then they created a sensorimotor coupling between two oscillator brain models. Testing the coupled oscillator brain models in different tasks of reciprocal social interaction, they found that the anatomical similarity between humans could explain a tendency to enter in synchronized brain activity, while being in the same context (coupled or not). Kelso et al. (2013) summarize their research into the oscillatory and bidirectional nature of brain dynamics, i. e., within the brain between neuronal ensembles, or between two brains in social interaction. They argue that there are basically two forces at play in dynamic coordinating brain systems: 1) The coupling between information exchanging systems allows to distribute information and by that allows for joint action. 2) The autonomy of individual systems controls the influence of other systems. The interplay between both, so they argue, leads to a form of self-organization. Or in other words: despite individual differences between neural networks in two interaction partner's brains, they will find a form of self-organization that forms or breaks patterns of harmonizing activity.

With a similar approach to modeling the dynamics of social interaction, Friston and Frith (2015b) tackle the problem of interpreting the intention behind communication. They create a model based on active inference, i. e., action fulfills our predictions about our own behavior. In coupling two models, with each producing sequences of bird songs, they found that the models could successfully predict the other model's bird song and attennuate any resulting prediction error. This showed that the hidden variable of the bird song must have been successfully inferred. Otherwise the song could not have been attennuated.

Although these models of interactive brain dynamics do give accounts of synchronized activity between brains, there are no explicit mental state attributions. These might not be necessary for coordinating behavior that leads to a form of belief coordination, in the form of synchronized activity. But to allow for the necessary reasoning and strategic coordination of belief attributions, a rule-based form of coordination recipes might be necessary.

3.6 DIFFERENTIATION AND CONTRIBUTION

The reviewed models all are missing one or more of the elements necessary for dynamically perceiving and production behavior during communication, to the end of coordinating inferred beliefs in direct social interaction. The necessary elements consist of 1) allowing for

dynamic and online perception and production of behavior, 2) enabling prior beliefs, biases or social norms to influence future behavior, 3) having beliefs about representations of behavior enable reasoning about other agent's mental states, and 4) applying processes of perception and action strategically to guide the dynamic coordination of such beliefs.

Some limitations of the presented models are simply due to their limited modeling scope. For example, models of motor coordination are simply not meant to incorporate a mentalizing perspective. But some of their limitations are more dependent on their modeling approach, rather than scope. One is the inability to adapt to unknown movements, i. e., learning is not supported or very limited. This can lead to limited robustness to infer the correct action understanding from real-world movement data. Indeed, a lack of robust action understanding directly influences the model's potential ability to infer mental states from an observed agent, rendering attempts of belief coordination impossible. Also, previous hierarchical models have not attempted to allow for simultaneous production and perception of behavior, as necessary to perceive another agent's behavior during one's own action production. In models of direct social perception, e. g., modeling brain dynamics, no explicit mental state attribution is made. At the moment, such models lack the necessary rule-based recipes for coordination. To allow a form of coordination, an ability to implement more complex feedback structures is necessary that would allow strategic alterations to an agent's own brain dynamics. But also, simple heuristics for mentalizing (as applied in Kahl and Kopp, 2015) might not be enough, as they are not able to extend to new situations or be robust enough to handle variable real-world input.

The cognitive modeling approach presented in this work attempts to tackle these limitations. We call it Hierarchical Predictive Belief Update (HPBU). HPBU is a hybrid model that combines a linear dynamic systems approach with a form of hierarchical and empirical Bayesian updating. The hierarchical levels of increasing abstractions allow to model sensorimotor processes as well as mentalizing processes. This is necessary for dynamic and online processing of behavior, inference of mental states, and the automatic or strategic coordination of beliefs during social interaction. The model makes similar assumptions as the Helmholtz machine, i. e., trying to find approximate representations for the statistical dynamics in the signal. Thus, it will learn representations that are grounded in the dynamics the system is exposed to, during action and perception. But instead of two seperate streams for recognition and for generation (or forward and inverse models, e. g., as in Wolpert and Kawato, 1998), only one generative model will be necessary to account for – and minimize – uncertainty during perception, while catering to the need for stabilization during action. As its foundation, with the modeling approach presented here, we argue

for the use of predictive processing, with the long-term minimization of uncertainty, or the so-called free energy principle. Through that, it will be able to handle uncertainty in the perceived behavior by allowing for uncertainty minimizing beliefs to become predictive in the hierarchy top-down, and have an effect on the belief coordination between interaction partners.

4

MODELING A PREDICTIVE PROCESSING HIERARCHY

The combined modeling approach, called HPBU (Hierarchical Predictive Belief Update), covers two parts related to the functional networks of the social brain: the sensorimotor part, and the mentalizing part (see ch. 5). First, we will visit the modeling of the predictive processing hierarchy (sec. 4.1). After that the computational sensorimotor processing part will be described (sec. 4.2). It is based on a hierarchy over increasingly abstract representations about behavior, which is learned using a self-supervised approach (sec. 4.2.5).

4.1 HIERARCHICAL PREDICTIVE BELIEF UPDATE

First we will visit the modeling assumptions, and the corpus data that will be used. Then, a detailed description of the actual model will follow, starting with a description of the generative model embedded in a dynamic environment. It is followed by a description of the information passing and belief updating between layers and with the environment.

4.1.1 *Modeling assumptions*

Computational cognitive modeling necessarily needs a set of assumptions that guides the kinds of algorithms that can be applied under limitations of cognitive plausibility. *Cognitive plausibility* is guided by results from cognitive psychology and cognitive neuropsychology, and optimally would also take resource limits into account, which would even limit the computational complexity. Included in this list of assumptions are also some which may have been included based on implementation necessity. I will try to mark these as such, since it is problematic to confuse assumptions driven by implementation concerns with those driven by theory (Cooper and Guest, 2014).

Assumptions that have guided the work on the HPBU, and on the model of the sensorimotor system specifically, are:

- First of all, we assume for the model that the cortex is hierarchically organized into functional processing entities, called cortical columns. This does not entail a singular hierarchy, rather it is more likely that there are multiple uni-sensory hierarchies that overlap with others in multi-sensory areas, which might further be processed in a hierarchical manner. Such an assumption entails an ontological structure, of primitive representations in

lower levels of the hierarchy that are processed to become ever more abstract representations, higher up in the hierarchy.

- Further, representations in the hierarchy are combining visual, motor, and proprioceptive aspects of action, if available. Consequently, they are used as high-level, or visuomotor representations of action and their outcomes. This also points to the embodied nature of cognition (Wilson, 2002). This assumption is based on converging evidence for the multimodal nature of representations that can be found in somatosensory, primary motor areas, and premotor areas of the human brain, which can code for both visual and proprioceptive information (Fogassi and Luppino, 2005; Gentile et al., 2015; Graziano et al., 2000; Pipereit et al., 2006; Wise et al., 1997).

- The hierarchy will exhibit a generative model, with each level consisting of a generative process. Their outputs are utilized as predictions. This is of importance, because when the difference between a predicted and perceived effect is minimal, so is the need for further processing of that sensory percept, which saves energy and aids in the survival of the organism. Also, given the high energy consumption of our brain, compared to other organs of our body (Raichle and Gusnard, 2002; Sokoloff et al., 1955), it is vital for us to conserve as much energy as possible.

- Representing the statistical dynamics in the environment is a goal similar to that of the Helmholtz machine, but instead of two seperate streams, for recognition and for generation (as discussed in par. 3), only one generative model will be necessary to account for – and minimize – uncertainty during perception, while catering to the need for stabilization during action. This also pertains to how uncertainty can be represented in the form of probabilities.

- Dealing with uncertainty is central to cognition. One account of the processes of cortical function that takes uncertainty into account in the form of probabilities, is *predictive processing* (Clark, 2016; Friston and Kiebel, 2009). This can be understood as a more general mechanistic property of efficient information processing systems that are in exchange with other social agents, and with their environment (Friston, 2013).

- One basic assumption that might be implementation driven, is that the dynamic environment unfolds as an ordered sequence of states, where input can induce dynamic trajectories in representational state space embedded in the brain. Further, it is assumed that the brain is a generative model of such trajectories, but in a hierarchical form that enables it to categorize and represent similar sequences. Such a hierarchical generative model can exhibit

multiple time scales, e. g., sequences covering longer times scale can generate sequences of shorter time scale. A similar idea was previously explored in Kiebel et al. (2009).

- Principles underlying the information exchange between levels of the model depend first and foremost on the hierarchical structure, with main connections to direct neighbours, as it was found between cortical columns for the supposed microcircuitry for predictive coding (Bastos et al., 2012). In addition, long-range connections transmit necessary information to levels when necessary, breaking up the strict hierarchical communication. This could be interpreted as an implementation level assumption, but as the human brain does not only consist of the neo-cortex, and has ample thalamo-cortical connections, so does HPBU not exclusively model intra-cortical processing.

- As already discussed (see sec. 3.5), the computational cognitive model presented here is a hybrid model combining a linear dynamic systems approach with an empirical and hierarchical Bayesian update, trying to find approximate representations for the statistical dynamics in the input signal. Thus, representations will be grounded in the dynamics the system is exposed to during action and perception.

- To represent the statistical dynamics in the environment, surprising events are learned by extending the model's state space. That is, the discrete distributions as well as the representations are extended to account for the surprising event on different levels of the hierarchy.

As described earlier (see sec. 2.1.3), a specific focus of the present research is non-verbal communication, such as writing, gestures or social gaze. Representations of speech and gesture have been found to be intimately connected (e. g., Alibali et al., 2001; Goldin-Meadow and Beilock, 2010), with gesture also being part of a loop of self-directed speech – similar to writing – that can help in your own thinking process (McNeill, 2008). Generelly, there is ample evidence for non-verbal communication to be an ontogenetic and phylogenetic precursor for verbal communication (Tomasello, 2008). In order to train and evaluate our computational model of non-verbal communication, we will now visit the handwriting corpus that was recorded for the purpose of studying this.

4.1.2 *The corpus of handwritten digits*

Usually when handwriting recognition is tested in the machine learning and artificial intelligence literature, the MNIST (or modified NIST) dataset comes to mind (Lecun et al., 1998). MNIST consists of 60.000

training images and 10.000 test images of size-normalized images of hand-written digits. MNIST was not used for training HPBU, because it does not contain the necessary sequential information.

Another dataset contender was released only a couple of years ago. The Omniglot corpus was first constructed to be tested on a novel one-shot Bayesian Program Learning approach by Lake et al. (2015). Omniglot contains the sequential and temporal information of multiple sequences of each 1623 hand-written character from 50 writing systems. The handwriting sequences were collected using the Amazon Mechanical Turk service, where participants had to draw the shown characters by hand. The drawings vary greatly in writing speeds, consistency, and fluency. This may be due to the artificial nature of the recordings, where participants had to write characters that they were not familiar with. To circumvent known problems with

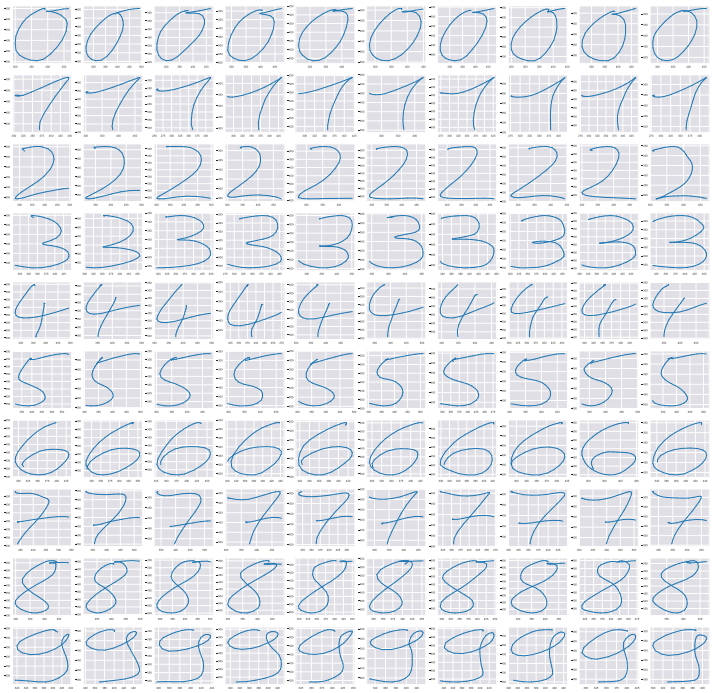


Figure 4.1: Examples of handwriting in the corpus of handwritten Hindu-Arabic numerals, from one recorded participant.

the existing datasets, we recorded our own, of all ten Hindu-Arabic numerals, written by 5 participants that were familiar with writing these on a daily basis. Each participant recorded each digit 10 times, resulting in 500 written digits (for an example by one participant, see fig. 4.1).

Recordings were collected using an iOS app that was first developed for the purpose of recording the corpus of handwriting digits on an 6th generation Apple iPad, using the Apple Pencil as a means of input. This setup was chosen for its natural input characteristics.

In the resulting 500 written digits a large amount of variance can still be found with respect to the speed and sequence of how digits are written in each individual's writing style. This certainly creates problems for algorithms to compare these sequences, because many are only able to compare sequences of similar size. We will return to this problem later. Still, one can argue that these sequences of handwriting are at least natural in their appearance and dynamics.

4.1.3 *Generative model and the environment*

We now dive into the formal description of the environment the model has access to and the model itself.

Each level L of the hierarchical generative model maps its *internal* state space L_i onto the domain of its next lower level L_e via *map*, representing the actual mapping function $map : L_i \to L_e$ of each level of the hierarchy. Conceptually, this describes the *top-down* influence from a higher level in the hierarchy. Also, the model maps from external to internal states (L_i) to minimize entropy. Conceptually, this describes the *bottom-up* influence from a lower level in the hierarchy. What is described as *external* depends on where in the hierarchy the level is situated.

Entropy minimization can be described as a function of $h : L_e \to L_i$ which maps external states onto internal states in a way that minimizes entropy (see eq.4.2).

$$h(L_e) = \arg\min_{l_i \in L_i} H(L_e|l_i) \tag{4.1}$$

$$H(L_e|l_i) = -\sum_{l_e \in L_e} P(l_e|l_i) \ln P(l_e|l_i). \tag{4.2}$$

The sensorimotor part of the hierarchy consists of four levels, the C level (Schemas), the S level (Sequences), and in the lowest levels, M level (Motor) and V level (Vision). For levels M and V the external states describe the sensory states of the system, which they will map to if possible.

The environmental state space consists of $X = \mathbb{R}^2$ and discrete time $T = \mathbb{R}$. It is defined in terms of a dynamical system of $(X, T, \varphi, \vartheta)$. $\varphi : T \to X$ is a function of discrete movements over time, observable by the system. $\vartheta : X \times T \to X'$ is a function of movements (from updating positions) in the environmental state space. The model must figuratively be understood as a box that connects to a hand that is glued to a writing pen. The hand can be moved through ϑ.

The specific levels of the generative model (C, S, M, V) are sequentially updated. They are updated in sequence, starting at the top, from it's next higher and next lower levels (if available). They thereby learn to represent and produce the states in their next lower level,

and observe joint positions in the environment $x = \varphi(t)$, which it can influence using a movement $m \in M$, resulting in $x' = \vartheta(m, t)$.

Each level of the generative model contains its own discrete probability distribution, where each probability describes a hypothesis of one discrete representation. The calligraphically styled level-variable denotes the random variable contained in that level, e. g., $\mathcal{C}$ is the random variable contained in level C. This results in a joint distribution of the hierarchy, as follows:

$$P(\mathcal{C}, \mathcal{S}, \mathcal{M}, \mathcal{V}, x) = P(\mathcal{C}) \cdot P(\mathcal{S}|\mathcal{C}) \cdot P(\mathcal{V}|\mathcal{S}) \cdot P(\mathcal{M}|\mathcal{S}) \cdot P(x|\mathcal{M}) \tag{4.3}$$

4.1.4 *Inter-level communication*

The here presented hierarchical generative model depends on a fast exchange of information between levels. As already mentioned, the generative model will be updated in sequence, starting at the level at the top of the hierarchy.

The difference between continuous and discrete states in the context of active inference is well discussed in Friston et al. (2017a). They explain that under discrete states the updating of beliefs depends on a so-called message passing scheme, while under continuous states we would describe a predictive coding scheme. The difference being that under predictive coding only prediction errors are communicated to a next-higher level in the hierarchy. The updating of beliefs, and how message passing is handled in the model of discrete states presented here will be discussed later (see par. 4.1.4).

Generally, there are three kinds of information being transmitted between levels in the hierarchy. Two of those are transmitted between direct neighbours, while one is used for either broadcasting or more specific transmission to other levels. First, information being transmitted to direct neighbouring levels are *prediction* data and *feedback* data. Prediction data consists of a level's posterior probability distribution, containing updated beliefs for all representation hypotheses of that level, being send to that level's next-lower neighbour. Second, the same data is being send to the next-higher neighbouring level as feedback information.

To clarify, this model does not implement a standard hierarchical Bayesian updating scheme. Of course, this is a probabilistic hierarchy akin to a hierarchical Bayesian model. But, a linear dynamic update at each level is a vital part of the belief updating that incorporates top-down and bottom-up information, and also, new representations can be added to the levels as needed. There are no standard Bayesian update approaches to updating these hierarchical Bayesian models in this dynamic sense. Rather, HPBU implements empirical Bayesian updating to incorporate predictive *top-down* information with the posterior from the last time step, which has become the prior for the current time step. Similarly, an empirical Bayesian update incorporates

bottom-up feedback information with the posterior from the last time step. This results in a posterior that is based on bottom-up information (P_{bu}), and a posterior that is based on top-down information (P_{td}). Combining these is up to a variational updating scheme based on Kalman filtering (which can be understood as a linear dynamic model), which will be discussed next.

The third kind of information transmission is described as *long-range connections*, which are used to broadcast information relevant for other levels. Long-range connections are also used to transmit information similar to the function of a *corollary discharge* or *efference copy* to inform a level about what information to expect. This kind of information transmission is vital for closing the visuo-motor coordination loop, as will be explained soon (see par. 4.2.2).

CALCULATING FREE ENERGY The HPBU model is defined as a hierarchical generative model which learns to predict and explain away prediction errors and in this sense minimize its free energy. The free energy, calculated for each level, is used as a measure of uncertainty of its fit to the environment. In HPBU this measure of uncertainty is used to calculate weighted belief updates, and to threshold the learning of new representations. Free energy describes the negative log model evidence of a generative model that tries to explain hidden states, e. g., events in the environment.

Evidence corresponds to probabilities of data from the environment, given the model at hand. Each level of the hierarchy contains a discrete random variable $\mathcal{X}$, for which two states are represented separately. The prior P stands for the top-down posterior P_{td}, while the posterior Q stands for the bottom-up posterior P_{bu} of $\mathcal{X}$. So in effect, free energy will be caculated between a variable that incorporates predictions from the next-higher level, and a variable that incorporates (external) sensory-, or feedback information from the next-lower level.

Surprise: each model's self-information which, when averaged over all states, corresponds to the model's entropy.

The free energy in each level is expressed as the sum of *surprise* over the level's internal prior state P, and a cross entropy of two states (the prior P, and a posterior Q after evidence has arrived). Each level's state-representation corresponds to a model (or policy) $\mathcal{X}_i \in \mathcal{X}$. Surprise describes each model's self-information, or negative log model evidence $-\ln P(\mathcal{X}_i)$ (Parr and Friston, 2019). Averaged over all models represented in $\mathcal{X}$, surprise corresponds to the entropy of the level's states, as in:

$$F(\mathcal{X}) = H(P(\mathcal{X})) + D_{KL}(P(\mathcal{X})||Q(\mathcal{X})) \tag{4.4}$$

$$= -\sum_i P(\mathcal{X}_i)\, \ln P(\mathcal{X}_i) + \sum_i P(\mathcal{X}_i) \cdot \ln \frac{P(\mathcal{X}_i)}{Q(\mathcal{X}_i)} \tag{4.5}$$

This formulation is a form of expected free energy, because here the prior probabilities of possible outcomes are not explicitly involved as they are part of the represented states (Parr and Friston, 2019).

In other words: in this hierarchical generative model, each level's state-representations correspond to models (or policies) that contain possible future outcomes.

Free energy in HPBU is calculated based on current information, to select a currently best fitting model. This best fit (or likelihood) is calculated differently, depending on the kind of representations a level contains, i. e., sequence-based or cluster-based. The next possible steps (toward an outcome) will be calculated in sequence-based levels, translating their sequence-based representations into possible observations of their next-lower level. In cluster-based levels the membership of a perceived sequence-representation is evaluated, to establish *cluster* or *schema*, in effect forming abstractions over similar sequences.

VARIATIONAL BELIEF UPDATE To update level beliefs, both posteriors will be combined to form the current level-posterior P_t. In this setting, the top-down posteriors play the role of prior beliefs, while the bottom-up posterior is the evidence. The difference between the bottom-up and top-down posteriors can be treated as a prediction error. Both are entered into a Kalman filter to create a linear dynamic system. The to-be calculated Kalman gain K is used to differentially weigh bottom-up evidence against top-down predictions, and plays the role of a precision-weighting. K is a function of free energy F (eq. 4.9) and the precision factor π, i. e., the inverse of the variance of the prediction error (eq. 4.8).

This is similar to the identified cortical microcircuitry for predictive coding (Bastos et al., 2012), where it was described that connections between cortical columns are mostly inhibitory. Here, belief updating is described as modeling a top-down inhibitory influence that attennuates the bottom-up information:

$$P_t = P_{td} + K_t(P_{bu} - P_{td}) \tag{4.6}$$

$$K_t = \frac{F_t}{F_t + \pi_t} \tag{4.7}$$

$$\pi_t = \ln \frac{1}{\sigma^2(P_{bu} - P_{td})} \tag{4.8}$$

$$F_t = H(P_{td}) + D_{KL}(P_{td} \| P_{bu}) \tag{4.9}$$

To be exact, the posterior probability distribution over represented beliefs are updated over the whole time course of perception or action. Thus, the belief update integrates the filter (eq. 4.6) into the *dynamically* updated hierarchical model context (eq. 4.7-4.9). Still, free energy describes the current upper bound on surprise (or model evidence), as a measure of uncertainty.

This belief update as well as the calculation of the Kalman gain are both performed with every new observation or prediction. If necessary, at every time step it allows for the selection of a new approximate maximum posterior representation, until free energy is

(hopefully) minimized after perception or action are finished. This is called *variational belief update*, performed in every level of the hierarchy.

Variational belief update: beliefs are updated dynamically while taking the current model uncertainty into account.

In the belief updating approach applied in HPBU, the difference between *perception* and *action* lies in the question: which of both drives the belief updates, the bottom-up or the top-down information? When it comes to calculating the level posterior distribution P_t, it is the Kalman gain K that sets the belief update towards either: maintaining the prior information from P_{td} to drive *action*, or towards having the prior being strongly updated by information from P_{bu}, to drive *perception*.

NORMALIZATION Before each belief update, the last level posterior P_{t-1} will be used as an empirical prior for calculating P_{bu} and P_{td}, respectively. We will soon discuss this for every level of the hierarchy.

P_t is normalized after each belief update. In this normalization step Laplace smoothing is used (Manning et al., 2008, pp. 193). It corrects for numerical errors that due to limits in floating point accuracy, can result in single probabilities reaching $P = 1$ or $P = 0$:

$$\hat{P}_t(\mathcal{X}_i) = \frac{P_t(\mathcal{X}_i) + \alpha}{\sum_i P_t(\mathcal{X}_i) + \alpha} \; \forall \mathcal{X}_i \in \mathcal{X} \tag{4.10}$$

Here, it simply adds $\alpha = 0.0001$ to each count, to leviate possible numerical errors. This smoothing step can be interpreted as a uniform prior, which is updated with the level posterior.

4.2 MODELING A SENSORIMOTOR SYSTEM

This section describes a predictive-processing based model of a sensorimotor system, which corresponds to the MNS. Later, the second part of HPBU will be described: the mentalizing part, which corresponds to the MENT (see ch. 5). Fig. 4.2 depicts a technical overview over the sensorimotor part of HPBU. It enables perception and production of behavior in a hierarchy of increasing abstractions over simple movement primitives, as represented in the lowest levels V (Vision) and M (Motor). Level S (Sequences) represents sequences of movement primitives and is part of the motor-coordination loop implementing active inference. The highest level of the sensorimotor part of the model consists of level C (Schemas), which forms clusters over similar level S representations.

4.2.1 *Level definitions and updates*

The model must figuratively be understood as a box that connects to a hand that is glued to a writing pen, which can be influenced, while the box also perceives the trajectories of what is written on a piece

Figure 4.3 has already been published in Kahl and Kopp (2018).

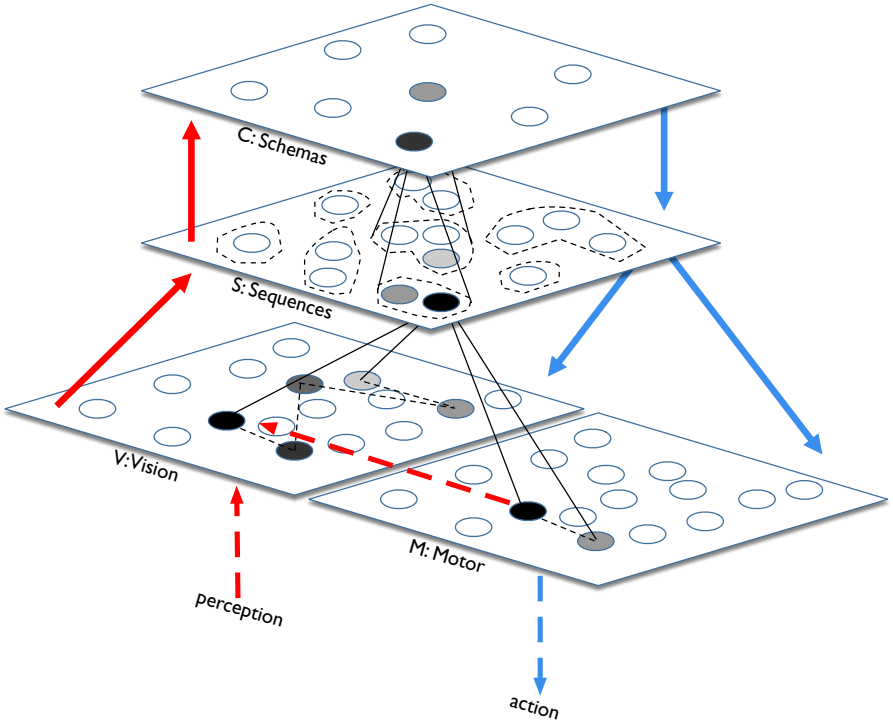


Figure 4.2: Depicted here is a technical overview of the sensorimotor part of HPBU, enabling the perception and production of behavior. At the top, level C (Schemas) clusters similar level S (Sequence) representations in the form of schemas. Sequence representations in level S allow the prediction of movement primitives in level V (Vision) and M (Motor). Producing an action sequence triggers the consecutive *active inference* of movement primitives in level M towards the predicted movement target from level S, thereby minimizing prediction error. Once minimized, level M will inform level V to check its success and close the motor loop.

of paper. Fig. 4.3 shows a sketch of the modeled cortical hierarchy of the sensorimotor processing part of HPBU and how it is connected to its environment. Predictions are sent top-down and compared with sensory (bottom-up) evidence to drive belief updates within the hierarchy.

At the top, in the C level abstract clusters of similar action sequences are represented. Below that, the S level represents sequences of visuo-motor acts. The lowest levels in the model hierarchy allow for action production, and proprioceptive feedback in M level and visual input and action feedback in the V. Red and blue lines represent bottom-up and top-down information propagation, respectively. The blue dotted line from V represents a visual prediction without any effect on the world, while the blue line from M has a causal effect. The red dotted line from M represents a long-range connection, further explained below.

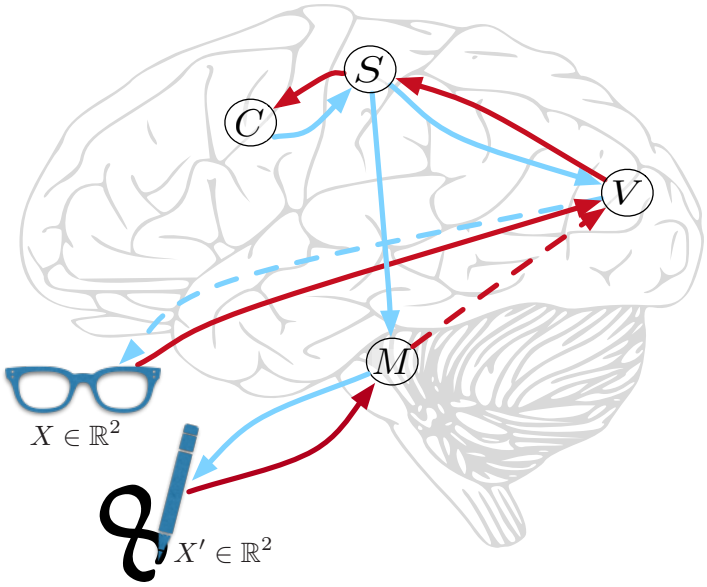


Figure 4.3: Depicted is a sketch of the modeled cortical hierarchy of the sensorimotor part of HPBU. Predictions are sent top-down and compared with sensory (bottom-up) evidence to drive belief updates within the hierarchy. Depicted levels have loose associations with the displayed cortical and subcortial structures. In schema (C) level, abstract clusters of similar action sequences are represented. Below that, the sequence (S) level represents sequences of visuomotor acts. The lowest levels allow for movement perception and production. Movements are updated $X' \in \mathbb{R}^2$ with proprioceptive feedback in motor control (M) level, and visual input $X \in \mathbb{R}^2$ in vision (V) level. Red and blue lines represent bottom-up and top-down information propagation, respectively. The blue dotted line from V represents a visual prediction without any effect on the world, while the blue line from M has a causal effect. The red dotted line from M represents a long-range connection.

Depicted levels have loose associations with the displayed cortical and subcortial structures. Schema and Sequence levels are associated with Primary Motor Cortex and Premotor Cortex areas which, as already discussed, are assumed to code for (visuo- or) senso-motoric forms of action sequences. Vision level representations code for the perception of movement directions, similar to area Medial Temporal (MT) in the visual cortex. The polar coordinates used in S are relative oculocentric coordinates (ϕ, r) of the visual field which, when seen in sequence, are similar to saccadic eye movements. Such a gaze-centered oculomotor frame of reference has been shown to code for the visual targets for reaching, and other actions (Ambrosini et al., 2012; Engel et al., 2002; Russo and Bruce, 1996). In Motor level M these coordinates will guide action in the form of movement goals. The production and proprioceptive feedback of movement in different directions is coded

in the Motor Control level M. It corresponds to reflex arcs, embedded in the tight coordination of basal ganglia, spinal cord, and cerebellum for the description of the motor coordination loop (see par. 4.2.2).

LEVELS AND REPRESENTATIONS At the top of the hierarchy we find level C (Schemas), which represents abstract clusters of similar action sequences in so-called schemas. Level C consists of cluster representations $\{c_1, \ldots, c_n\}$, and contains a discrete probability distribution $\mathcal{C}$ over these n discrete states. The calligraphically styled level-variable denotes the random variable contained in that level, e. g., $\mathcal{C}$ is the random variable contained in level C. Every discrete state's probability will be represented in its level's respective discrete probability distribution, e. g., for $c_i \in C$: $P(c_i)$ represents the respective cluster representation's probability from $\mathcal{C}$. Each schema c_n clusters sequences $S' \subseteq S$ by similarity, and finds a median in the cluster, representing the schema as a cluster prototype $\tilde{c}_n$. It maps to its next lower level S with $s : C \mapsto S$.

Below level C, sequence level S represents action sequences $\{s_1, \ldots, s_m\}$, and contains a discrete probability distribution $\mathcal{S}$ over these m discrete states. Each action sequence s_m contains a tuple of observed movements $(o_1, \ldots, o_k)$ in polar coordinates at time $t \in T$ with $o_k = (\theta, r)$, and the time delay between observations $(\Delta_2, \ldots, \Delta_k)$, with $\Delta_k = t_k - t_{k-1}$.

That is, S (Sequences) maps to it's next lower level V (Vision) with $\nu : S \mapsto V$ where S consists of sequences of *salient* states from V. To detect a salient event o_k in V the model free energy $F_t(V)$ is calculated (see par. 4.1.4), given two consecutive input events from environmental state space $\varphi(t)$. If then $F_t(V) > F_{t-1}(V)$, a salient event was detected, as the updated model was not able to correctly predict the current input event.

The lowest levels of the hierarchy represent visuomotor primitives, allowing for action production, visual input, and proprioceptive feedback. V (Vision) and M (Motor) code relative movement angles of a joint in $i = 16$ directions $\{\nu_1, \ldots, \nu_i\}$. Visually, this is represented in V, which also contains a discrete probability distribution $\mathcal{V}$ of these i discrete states. Each ν_i represents a movement visually perceivable by the model. In addition, a jump in writing (placing the pen at another point to resume writing) is detected and stored as a jump-flag. Relative movement angles are also coded in M, where each m_i is a possible movement applicable to the joint, which can then be proprioceptively perceived by the model. M also contains a discrete probability distribution $\mathcal{M}$ of these i discrete states. In addition, level *M* is the only level that can influence the environmental state space X using:

$$m' = \underset{m_i \in \mathcal{M}}{\arg\max}\, P(m_i), \tag{4.11}$$

which results in a movement in $\vartheta(m', t)$.

With this in mind, let us turn to each level's update mechanism of both: the bottom-up posterior and the top-down posterior.

INTRA-LEVEL BAYESIAN UPDATE The presented model (so far) consists of just three different update mechanisms. Later in the section on the extended version of HPBU that covers the mentalizing part, we will revisit the update mechanisms, where necessary.

The type of level, either sequence-based or cluster-based, will determine the kind of update. With the exception of the Vision and Motor Control levels, two types of level-updates are differentiated based on the kind of representation: a sequence-based update and a cluster-based update.

For a technical overview of these and additional calculations embedded in the input and output of levels S and C, please have a look at fig. 4.4.

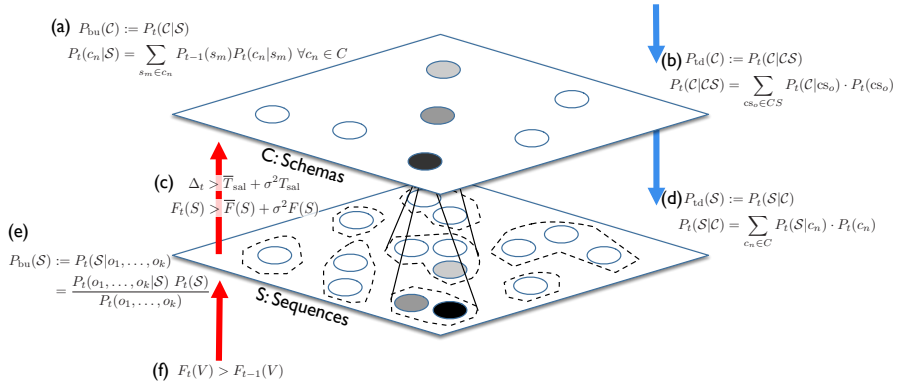


Figure 4.4: This technical overview shows the following input and output calculations embedded in levels C and S: (a) the bottom-up update of $\mathcal{C}$ is described in eq. 4.13, (c) describes the detection of suprising sequences (par. 4.2.5), (d) is the top-down (eq. 4.15), and (e) the bottom-up update of $\mathcal{S}$ (eq. 4.16), with (f) describing the detection of surprising movements (par. 4.2.3).

The schema level C clusters representations from sequence level S by similarity, to find clusters of similar action sequences, along with a median action sequence that represents the cluster as a prototype.

Comparing action sequences will be discussed later (see par. 4.2.3) as well as a detailed explanation for the clustering algorithm used here (see par. 4.2.5).

Level C calculates it's bottom-up posterior $P_{bu}(\mathcal{C})$ using the soft evidence "all things considered" method over updated sequence probabilities, given the schema *cluster* it belongs to (for more information, please see Darwiche, 2009, ch. 3.6.1).

$$P_{bu}(\mathcal{C}) := P_t(\mathcal{C}|\mathcal{S}) \tag{4.12}$$

$$P_t(c_n|\mathcal{S}) = \sum_{s_m \in c_n} P_{t-1}(s_m) P_t(c_n|s_m)\ \forall c_n \in C \tag{4.13}$$

The top-down posterior for level C will be discussed later (see sec. 5.2), when there is a next-higher level to update from, in the description of the mentalizing model.

Level S calculates its top-down posterior $P_{td}(\mathcal{S})$ from a mixture of experts.

$$P_{td}(\mathcal{S}) := P_t(\mathcal{S}|\mathcal{C}) \tag{4.14}$$

$$P_t(\mathcal{S}|\mathcal{C}) = \sum_{c_n \in C} P_t(\mathcal{S}|c_n) \cdot P_t(c_n) \tag{4.15}$$

For the bottom-up posterior $P_{bu}(\mathcal{S})$ the sequence probability needs to be calculated, given observations from V, which represents only singular observations of movement. So in order to calculate their likelihood, observations are collected in a temporary sequence that grows over time $s' = (o'_1, \ldots, o'_k)$. The likelihood for the temporary sequence, given each known sequence $P(s'|s_m)$ is the sequence difference dtw (see eq. 4.28). In addition, the sequence difference is weighted by an exponential factor, which calculates the temporal precision of the observed state (see par. 4.2.3). This is in effect comparable to calculating a joint probability for all observation events $P(o'_1, \ldots, o'_k|s_m)$. Calculating the posterior then simply is a Bayesian inversion:

$$P_{bu}(\mathcal{S}) := P_t(\mathcal{S}|o_1, \ldots, o_k) = \frac{P_t(o_1, \ldots, o_k|\mathcal{S})\ P_t(\mathcal{S})}{P_t(o_1, \ldots, o_k)}. \tag{4.16}$$

As a prediction for lower levels of the hierarchy, i. e., levels V and M, a complete sequence would be of no use. The sequence would not map to the representations associated with the discrete probability distributions in these levels, i. e., relative movement angles of a joint in 16 directions. Thus, we need to obtain the probability of all possible next observations ($\forall v_i \in V$), with the prior observations $o'_1, \ldots, o'_k$, given the predicted sequence s_m:

$$P_{td}(V) := P_t(\mathcal{V}|\mathcal{S}) \tag{4.17}$$

$$P_t(\mathcal{V}|\mathcal{S}) \approx P_t(v_i|o'_1, ..., o'_k, s_m) = \frac{P_t(o'_1, \ldots, o'_k, v_i|s_m)}{P_t(o'_1, \ldots, o'_k|s_m)} \tag{4.18}$$

The resulting distribution is compatible with representations of M level and is used similarly as $P_{td}(\mathcal{M}) := P_t(\mathcal{M}|\mathcal{S})$.

The bottom-up posterior $P_{bu}(\mathcal{V})$ is a mapping from environmental state space $\varphi(t)$ to the model's movement reportoire, using a parameterized gaussian likelihood function for each available movement $v_i \in V$, given $\sigma = 0.1$.

$$P_{bu}(\mathcal{V}) := P_t(\mathcal{V}|\varphi(t)) \propto P_t(\varphi(t)|\mathcal{V})\ P_t(\mathcal{V}) \tag{4.19}$$

$$P_t(\varphi(t)|v_i) = e^{-\frac{(\varphi(t)-v_i)^2}{2\sigma^2}} \tag{4.20}$$

The same applies also for level M. For a formal summary of these update mechanisms for the level S to levels V and M complex, have a look at fig. 4.5.

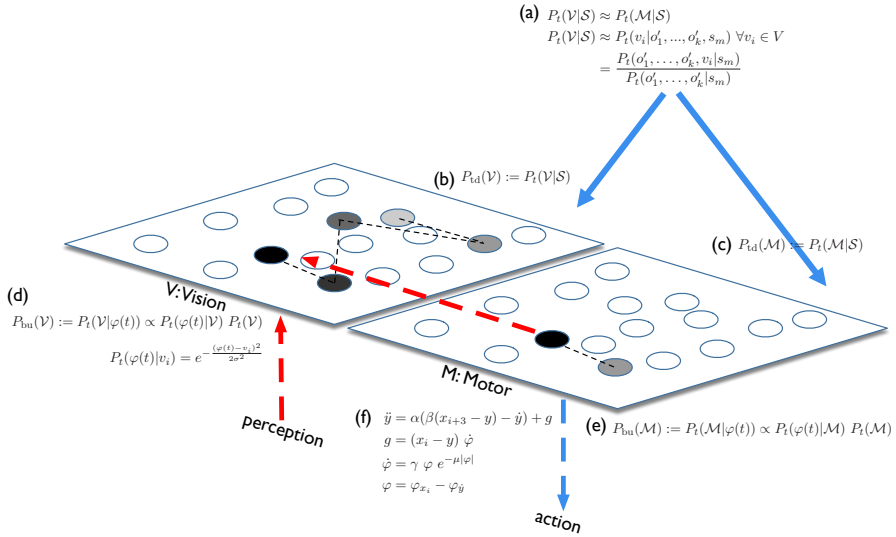


Figure 4.5: This technical overview shows the following input and output calculations embedded in levels V and M: (a) is the result of a sequence prediction, described in eq. 4.18, (b) shows how this sequence prediction becomes the top-down posterior of $\mathcal{V}$, which is approximately similar to the update of $\mathcal{M}$ in (c). (d) is the bottom-up update of $\mathcal{V}$ (eq. 4.20), which is also similar to the update of $\mathcal{M}$, while (f) describes the damped spring system for movement generation (eq. 4.24).

4.2.2 *A model of active inference*

Let us now turn to the other side of the coin, i. e., how the described representations and updates can lead to action, performed in the environment. The model of active inference will be described, continuing the discussion for a possible solution to the problem of motor coordination.

MOTOR COORDINATION In the sensorimotor hierarchy the lowest levels are the ones responsible for motor coordination. Specifically, the V and M levels represent two aspects in active inference that are necessary for motor coordination. As described in the previous subsection, V receives coordinates of a writing trajectory at discrete points in time. So at each point in time, writing is represented in the form of a discrete probability distribution, over a discrete set of writing angles.

Friston (2011) argues that the detailed planning of movements should be replaced by a free-energy minimizing application of reflex arcs. In HPBU we apply the oculocentric information, stored in the sequence representations in level S, to guide action, and in the process

Parts of this section on the motor coordination in the sensorimotor hierarchy has previously been published in Kahl and Kopp, 2018.

circumvent the need for detailed programming of motor commands. The relative polar coordinates are send to level M, where they act as action targets. In level M, a reflex arc in the form of a damped spring system realizes the motion toward the action target. The action target defines the spring's point of equalibrium at the relative polar coordinate, so that the movement realization just has to follow simple equations of motion.

This implementation of active inference is formally related to the equilibrium point hypothesis (Feldman and Levin, 1995). In other words, information from level S contains top-down predictions of the proprioceptive consequences of movement, similar to the argument for motor commands also just being predictions (Adams et al., 2012). They are regarded as setting and equilibrium or set point to which the motor plant (the moving joints) converges, via the engagement of motor reflexes. Similar to the ideomotor principle, once the movement goal is set, the motor system will select and apply the movements necessary to reach it (Prinz, 1990).

To allow for smooth and curving trajectories that are similar to handwriting in spatial and temporal properties, Dynamic Movement Primitives (DMP)s have been considered. DMPs have been used for modeling attractor behaviors of autonomous nonlinear dynamical systems with the help of statistical learning techniques (Ijspeert et al., 2013). Here, their ability to learn and reproduce trajectories is not used. Rather, the damped spring system is configured similarly to a DMP. Instead of applying a forcing term f that activates the system's nonlinear dynamics over time, here an obstacle avoidance technique is used (as described in Hoffmann et al., 2009). Its force was adopted and inverted to actually move *towards* the goal in a goal-forcing function g (see eq. 4.24). The reason for this is that when we would now simply apply the spring system to each goal sequentially, it would accelerate toward and slow down at the goal. Several simulations showed that to keep up the momentum we need to look ahead several goals x_{i+3} (here 3 steps ahead) in the core spring system. The goal-forcing function will still visit each goal x_i sequentially.

α, β, γ and μ are constants that specify the behavior of the system. φ is the angle to the goal and y is the current position.

$$\ddot{y} = \alpha(\beta(x_{i+3} - y) - \dot{y}) + g \tag{4.21}$$

$$g = (x_i - y)\ \dot{\varphi} \tag{4.22}$$

$$\dot{\varphi} = \gamma\ \varphi\ e^{-\mu|\varphi|} \tag{4.23}$$

$$\varphi = \varphi_{x_i} - \varphi_{\dot{y}} \tag{4.24}$$

The resulting acceleration $\ddot{y}$ will be twice integrated, before it is applied as an environmental state space position $\vartheta(y, t)$, with $t \in T$ and $y \in X$ (see definition of the dynamical system in par. 4.1.3).

Free energy in level M is minimized by moving towards the target in the predicted manner, i. e., with the optimal relative movement direction. This allows an action to be completed, by converging towards the spring system's equilibrium point as the target. But how can the visual system be informed, getting its information from level V, in order to close the visuo-motoric loop?

CLOSING THE MOTOR LOOP A very important aspect of the presented approach to active inference is a missing feedback connection between levels M and S (visible as a stroked, red arrow in fig. 4.3 and fig. 4.5). Once level M reaches the subgoal of an action sequence a signal is directly send to level V via a long-range connection. It informs level V of the location where the joint should have been moved, so it can evaluate whether the simultaneously received visual information can confirm the proprioceptive information. This closes the motor coordination loop.

The missing feedback connection between levels M and S has two reasons:

First, in order to investigate if motor coordination can rely on visual information alone to drive motor coordination, it is getting only sporadic proprioceptive feedback from level M to evaluate. This is different to the approach by Friston (2011), which relies heavily on proprioceptive information to make reflex arcs conform to action predictions through active inference. HPBU's active inference model spans a wider motor coordination loop and only loosly constraints motor control. This allows for variability in redundant aspects of the movement (similar to Todorov and Jordan, 2002), to reliably reach the next target (as briefly discussed in par. 3.5.2). Redundant aspects of movement are, e. g., the set of all possible arm configurations that allow the fingertip to reach to movement target.

Second, by making the model's sequence coordination independent from direct proprioceptive feedback, it allows for future developments that might enable it to associate actions in the world with intended effects that do not have to influence the motor system directly. An example might be the more distal action effect of pressing on a switch to turn on a light.

MOTOR CONTROL SUMMARY When action production is initiated, the motor coordination loop starts in the sequence layer, where an action sequence is selected. An action contains a sequence of movement primitives, consisting of tuples of oculocentric polar coordinates, and relative timing information. A movement that realizes the tuple information actually defines the movement target along with information about the predicted movement duration. The next movement primitive in sequence will be communicated to levels V and M. V receives the primitive in a form of an efference copy, for maintenance

of the correct priors for later comparison (see par. 4.2.4. M receives the primitive for realization through movement.

Simultaneously, the realized movement is visually perceived, while the action is being produced. This way the produced action sequence can be evaluated immediately, by comparing the perceived with the produced action. In addition, when level M believes that it has reached its target, it will signal level V to check for deviations between intended and perceived location of the moved joint, using a long-range connection.

As long as the predicted action sequence is successfully evaluated, each movement primitive will be sent for production one by one, until finished.

4.2.3 *Handling action sequences*

The comparison of action sequences is vital for evaluating the correctness of action predictions, during mere perception of other's actions, and also during production of own actions.

SEQUENCES OF SURPRISING MOVEMENTS The goal of the overall model is to minimize free energy by predicting and correcting for the statistical irregularities in the signal. The statistical irregularities that HPBU has to deal with are deviation from previous movement directions. Such statistical irregularities are informative, because they deviate from the previously predicted movement, and lead to prediction errors. This is reflected by increases in free energy and sometimes leads to model-switching, i. e., the prediction of a better-fitting representation of future movement. The detection of movement deviations as surprising events, are errors in the predictability of the signal, and can be used to find segmentation boundaries, that structure the signal in an information-theoretic perspective. This idea was developed by Zacks et al. (2007) into what they call Event Segmentation Theory.

As discussed, we confront the model with data from a previously collected corpus of handwritten digits. The drawing strokes contain dynamic movement information. Level V detects surprising strokes in the sudden increase of free energy, with respect to the last time step, as:

$$F_t(V) > F_{t-1}(V) \tag{4.25}$$

Information about this surprising stroke consists of the writing and its length, which are both transformed into the oculocentric reference frame. Also, the jump-flag is stored. More specifically, the relative coordinates are transformed into a relative polar coordinate, with the last surprising stroke coordinate at its center. Such a gaze-centered oculomotor frame of reference has been shown to code for the visual

targets for reaching, and other actions (Ambrosini et al., 2012; Engel et al., 2002; Russo and Bruce, 1996).

This information – along with the amount of time passed since the last surprising event – is sent to level S, which stores sequences of such surprising events in a temporarily collected sequence s'. The oculocentric reference frame, used for storing surprising events, is also used for the generation of action.

COMPARING ACTION SEQUENCES Multiple approaches to compare sequences have been applied, such as the Riemann Distance measure, comparing temporal sequences of probability distributions, in the form of a Riemann manifold (not further discussed here). Also, the alphabetic Jensen-Shannon distance measure (Mateos et al., 2017) has been tested, which created an alphabet of repeating subsequences for comparing their occurrence frequency, in a given sequence. I will not go into detail on these approaches. Instead, I will describe the comparison algorithm that was actually applied in the model, i. e., the Dynamic Time Warping (DTW) distance measure, a computationally light-weight sequence comparison method. As we discussed when introducing the corpus of handwriting of digits, sometimes sequence comparisons need to be able to compare sequences of different lenghts, while concentrating on aspects of similarity without ignoring the temporal dynamics inherent in hand writing.

In a dynamic programming approach, the DTW measure evaluates the difference between temporal sequences by finding the overlapping subsequence with the minimally necessary temporal edits between them. Each edit has a cost associated to it (as described in eq. 4.28). Such costs include insertions and deletions of temporal steps so that alignment can again be preserved (e. g., Salvador et al., 2007). Here, the distance measure d was specifically chosen to have a minimum sequence distance return approximately 1, i. e., the parameterized gaussian distance from a perfect match ($\mu = 0$) with σ^2 being chosen accordingly. Using dynamic programming, the algorithm finds the path with the minimal amount of edits between all movement primitives i of sequence a compared with movement primitives j of sequence b, as in,

$$\mathrm{dtw}(a,b) = e^{-\frac{(\mathrm{dtw}_{\min}-\mu)^2}{2\ \sigma^2}} \tag{4.26}$$

$$\mathrm{dtw}_{\min} = \arg\min\left(\mathrm{dtw}_{a,b}(i,j)\right) \tag{4.27}$$

$$\mathrm{dtw}_{a,b}(i,j) = d(a_i,b_j) + \min\begin{cases} \mathrm{dtw}_{a,b}(i-1,j)+1 & \text{del} \\ \mathrm{dtw}_{a,b}(i,j-1)+1 & \text{ins} \\ \mathrm{dtw}_{a,b}(i-1,j-1) & \text{match.} \end{cases} \tag{4.28}$$

The first minimum is akin to a deletion (or *del*, from sequence *a* to *b*), the second minimum corresponds to an insertion (*ins*), and the third corresponds to a match or mismatch (*match*). This depends on function $d(a_i, b_j)$, which optimally would add zero if movement primitives are identical. A comparison between movement primitives which, here is coded in polar coordinates (relative to the last movement primitive), translates the angular-difference coding from $\theta \in [-\pi, \pi]$ into $\theta' = \frac{\theta+\pi}{2}$. This allows for a cosine difference for the angle, and a log difference for the radius r, both weighted by w_d, as in,

$$d(a_i, b_j) = \begin{cases} s_p & \text{jump}_i \neq \text{jump}_j \\ w_d \ (1 - \cos(|\theta'_i - \theta'_j|) + \frac{\log(|r_i - r_j|+1)}{s_w}) & \text{otherwise.} \end{cases} \tag{4.29}$$

This comparison also takes drawing jumps into account, with a penalty variable s_p for disaligned jump-flags (necessary in writing, e. g., a seven, five, or four). The log difference for the radius of the polar coordinate is also weighted using the variable s_w. Those variables as well as σ for the gaussian distance have to be chosen carefully to reach a good classification performance. For the evaluation described later in this work, the variables have been set to $w_d = 10$, $s_p = 15$, $s_w = 5$, and $\sigma = 100$, to create a balanced comparison.

In order to compare a temporarily collected sequence s' to all known sequences, the distance is calculated for every $s_m \in S$. Also taken into account is the temporal distance between the predicted and perceived delay until the action's consequence:

$$P(s'|s_m) = dtw(s', s_m) \cdot e^{-\frac{(\Delta t_{s'} - \Delta t_{s_m})^2}{2\ \pi_S^2}}. \tag{4.30}$$

The temporal distance between predicted and perceived delay is a parameterized gaussian with a decreasing variance, depending on the sequence level's precision π_S. In level S, as in all levels of the hierarchy, such a likelihood will inform the new level posterior, after belief updating (see par. 4.1.4).

This is how HPBU compares action sequences and estimates their likelihood, given known sequences and observations.

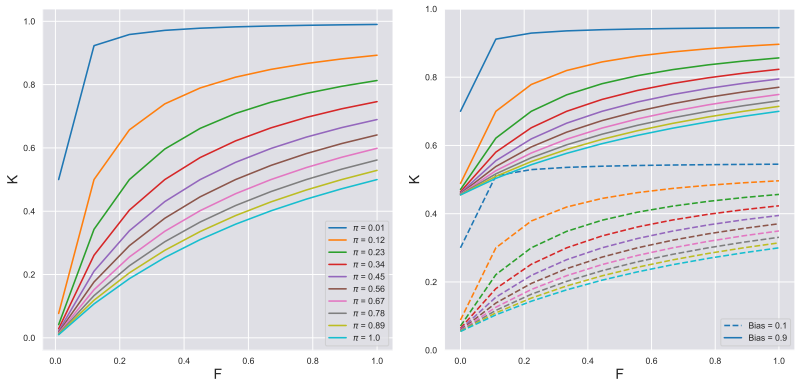
4.2.4 *Strategic action and perception*

During all interaction with the world our sensorimotor system needs to handle its uncertainty about the world, by balancing the influence of new evidence and make use of it, in order to meet its interaction goals. This need for balance when facing uncertainty entails a strategic component of deciding when evidence should be ignored or when

special attention is needed. Now we will consider the problem of strategically configuring the sensorimotor part of HPBU for action or perception. This also covers the problem of maintaining focus during action production. A focus, which is necessary because of the uncertainty in the signal, even of self-produced behavior.

BIASING KALMAN GAIN FOR PERCEPTION AND ACTION The sensorimotor hierarchy can be configured for action or perception using a biased K. First of all, this bias b toward either maintaining or changing prior information is driven by a level's uncertainty with respect to the external information. This is represented both: by the variance in the level's hypothesis space (reflected in the precision value π), and by the success of higher level predictions to attenuate prediction errors (as measured in the free energy F).

The resulting Kalman gain from different combinations can be seen in fig. 4.6 *(a)*. A high π results in a very slow response in K with increasing F, which mostly preserves the top-down prior, up until a balanced influence of prior and prediction error when both π and F are high. A low π results in a very steep response in K with even slightest increases in F, which leads to strong influences from prediction error.



(a) Kalman gain (b) Biased Kalman gain

Figure 4.6: (a) A high π results in a very slow response in K with increasing F, which mostly preserves the prior, up until a balanced influence of prior and prediction error when both π and F are high. A low π results in a very steep response in K with even slightest increases in F, which leads to strong influences from prediction error. (b) As an example of biasing Kalman gain, the K responses from (a) are biased either toward perception with $b = 0.9$, or toward action with $b = 0.1$.

For the strategic application of perception and action, two observations are important: for perception, a high K is necessary to stabilize the detected prediction error, in order to give it a chance to drive the belief update in higher levels by finding better hypotheses that again, minimize free energy. During action the opposite is true, as

there, strong prediction error would – in the worst case – overwrite an intended action sequence. To allow for stable action production, the top-down prediction has to be maintained, i. e., the influence of prediction error needs to be small, using a small K, setting the belief update towards maintaining P_{td}. Still, its influence needs to be high enough that it does not forfit all chances of allowing for prediction error to change perception.

In order to bias belief updates more strategically towards driving updates, either for perception or for action, we introduce a gain bias *b*. During perception, K is to be biased toward a higher gain, while for action K is biased toward a lower gain, still allowing for the uncertainty driven fluctuations.

Similar to a Kalman filter, the strategic perception-action bias *b* influences the update of K. It can be higher for perception (e. g., $b = 0.9$, to bias updating towards new information) than for action (e. g., $b = 0.1$ which would bias updating toward the prior),

$$K'_t = K_t + \frac{1}{2}(b - K_t) \tag{4.31}$$

resulting in a strategically biased Kalman gain K'_t. Thus, either P_{td} or P_{bu} becomes the driving signal for belief updates. To see its influence on the Kalman gain bias K, please see fig. 4.6 *(b)*. This bias parameter will become of vital importance when we come to the balance of attending to an interaction partner's behavior during social interaction.

MAINTAINING FOCUS: THE INTENTION SIGNAL As indicated previously, in addition to level M, level V also receives predictions of expected movement primitives during action production, in the form of an efference copy. *Corollary discharge* or *efference copy*, are motor command signals in the nervous system that send additional activity to distal areas for further processing. Such information is sent to brain regions that use sensory input, e. g., for comparison with visual or proprioceptive feedback, as suggested in the comparator for motor coordination (Wolpert et al., 1995). In the visual system, such information has been hypothesized to produce stable visual percepts, instead of eye-movement induced jumpyness (Sommer and Wurtz, 2008). Although, the concept of specific motor commands has been called into question, as in active inference, such signals would also be predictions about future actions (Adams et al., 2012).

Corollary discharge: a signal that in the nervous system is sent to additional distal areas, and is used for future processing, e. g., comparisons.

Similarly, own simulations (not shown) have convinced us that without an efference copy that acts as a stabilizing mechanism, action production in the model would suffer from visually induced jumpyness between behavior-producing models. This is due to the continuously ongoing perceptual inference process that compares the currently perceived action sequence with known sequences. In itself, it can – but will not necessarily – infer the exact same action sequence that is currently being produced. If action production would solely

rely on the maximum posterior action sequence hypothesis being selected for production, this would make ongoing production highly vulnerable to noise in the exterior action-perception loop. Such noise can take many forms, but the main source of noise, in the scenario presented here, is due to the dynamics of motor coordination of the damped spring-system based behavior, converging towards its movement goal. In other words: the dynamics of the spring system can create a discrepancy between the dynamics learned from the corpus of hand-written digits, and the dynamics of HPBU's motor coordination.

Intention signal: a corollary discharge information, used to maintain focus on the currently produced action sequence.

To counter-act visually induced jumpyness of action production, the described corollary discharge of motor command signals is interpreted as an *intention signal*. It maintains the currently selected action sequence throughout the production process on every level involved in visuo-motor processing For example, an intended sequence from level S would be $s_I \in S$, a schema from level C would be $c_I \in C$.

Generally speaking, there are two ways of chosing the intended action sequence in HPBU: One is to simply select the intended cluster's prototype $s_I = \tilde{c_I}^d \in c_I$, which will become necessary for strategies of efficient communication, as will be discussed much later. The other is to randomly sample a member sequence from the intended cluster. Such a random selection of a sequence, in order to produce an intended cluster, is the normal mode of operation for HPBU, unless otherwise stated.

The intention signal helps to strategically reduce the influence of bottom-up information influencing the top-down prior action prediction. This way, the perceptual inference process, comparing predicted with actual movements, can be biased towards an increased robustness of action production in light of unexpected movement dynamics.

The need for maintenance of action goals during action production points towards a dependence on robust representations, and shifts the computational burden to the acquisition and selection of the right prior, in the first place. Representations in the hierarchy take the role of priors that allow for predictable movement dynamics, and hence, can tune the generative model itself. To that end, we will now discuss the learning of motor sequences and schemas.

4.2.5 *Self-supervised learning*

As of yet, it was left out how the model collects, or "learns" representations of sequences, in the first place. These representations are the hierarchical model's primary source of priors to account for the sensory input, and to minimize free energy. It is then the success of these priors to attennuate prediction-errors that is evaluated, and which guides the model's reactions. These reactions have the form of belief updates that take into account the model's uncertainties through

precision weighting. Or in other words, the model picks up and represents the information-theoretic irregularities in the interchange with its environment to be able to account for them in the future.

HPBU employs a self-supervised learning process for self-organization, picking up these information-theoretic irregularities to form representations through uncertainty-driven action sequence selection, and similarity-based clustering. Already learned representations will act as temporary labels – or priors – to guide the prediction of perceivable actions, so that these known representations can be ignored, while unknown actions will be learned. This is a self-supervised approach, which similarly to the Helmholtz-machine (Dayan et al., 1995; Kawato et al., 1993), tries to find approximate representations for the statistical dynamics in the signal. This is different from supervised learning, where labels are given explicitly, or unsupervised learning, where no labels are given.

SEQUENCE LEARNING In the discussion about the uncertainty of sensory information and how to represent these, it is important to understand that, taken seriously, there is no such thing as a clear-cut differentiation between stimulus and response. A stimulus can be the consequence of an earlier interaction with the world, while even the preparation to react influences the perception of a stimulus. Also, during action execution, for the most time, we ignore that we perceive our actions, although even reflex arcs can influence further processing (Jordan, 1998). *"Thus, as the continuous dynamic closed loop of sensory input and motor output makes infeasible a true discrimination of stimulus from response, so does the embedded continuous dynamic closed loop of perceptual processing and action preparation make infeasible a true discrimination of perception from action [...]"* (Spivey, 2008, pp. 48).

So how can we actually discriminate between perceptual sequences? As we have seen in the description of the vision level V of the model hierarchy, surprising movement deviations can be detected as errors in the predictability of the signal. These events can be used to find segmentation boundaries that structure the signal in an information-theoretic perspective. This idea was developed by Zacks et al. (2007) into what they call Event Segmentation Theory (which we will discuss later in par. 3.3.1). This was applied in an account by Gumbsch et al. (2017), who propose to make use of the surprise that can be detected in transient free energy at event boundaries, to learn new action representations.

As discussed, the vision level of HPBU filters for salient movement primitives, which are communicated to the sequence level S. With every new salient movement level S extends a temporary sequence s', which is compared with already known sequences. In that process, level S detects surprising deviations from known sequences in the form of a sudden increase in free energy. To be invariant against small

fluctuations, the current free energy $F_t(S)$ is compared to a running average transient free energy $\overline{F}(S)$ and its variance σ^2, to signify a highly surprising deviation from known sequences if

$$F_t(S) > \overline{F}(S) + \sigma^2(F(S)). \tag{4.32}$$

In addition, not only the information-theoretic irregularity is – by definition – informative, but also temporal irregularities are. Detecting them is done similarly as for detecting surprising jumps in free energy. From the beginning of collecting salient movements in the temporary sequence, also a temporary sequence of *temporal* delays between salient movements is stored as $T_{sal} = [\Delta_2, \ldots, \Delta_{t-1}]$. Detecting salient temporal delays is then a matter of comparing the current delay Δ_t with mean transient delays and their variance, such that:

$$\Delta_t > \overline{T_{sal}} + \sigma^2(T_{sal}) \tag{4.33}$$

Those are two methods that allow for the detection of a salient segment – one information-theoretic and one temporal. The surprising temporary sequence s' will then be added to the list of representations, and will become represented in the discrete probability distribution $\mathcal{S}$ of level S. When only a temporally salient delay was detected s' will just be emptied, and all variables contingent on processing ongoing movement will be reset. This way, only movements that cannot be accounted for by known sequences will be added to the list of representations.

Adding a new action sequence representation effectively extends the discrete probability distribution, seemingly making updates incompatible. The extension takes place after the level posterior $P_t(\mathcal{S})$ was updated, so that before the next update cycle, the posterior can be renormalized (par. 4.1.4). This way, compatibility is restored. Also, bottom-up and top-down posteriors are calculated with $P_t(\mathcal{S})$ as a prior, before any other update.

CLUSTERING New movement sequences, when added to the list of representations, will in effect contribute to the whole model's ability to attennuate prediction errors from movement deviations. Also, a new level S representation extends its discrete probability distribution. This would make future updates incompatible, so in order to reestablish compatibility the sequence also has to be incorporated into the clusters of level C. This is necessary so that level C representations can be updated from, and predict, the complete level S state space.

Even though each movement consists of very individual movement dynamics, there are similarities between them. Especially task related movements or informative gestures result from underlying representations that have been triggered by a task or by a communicative intent.

Level C clusters movement sequences by their similarity to determine the hidden meaning behind the clustered similarity, i. e., the common hidden or latent variable, representing the clustered movement sequences.

In the current scenario, in which HPBU encounters only hand-written digits, a clustering approach with a predefined number of clusters would naturally suffice. For example, *k-means clustering* can be applied with a predefined number of clusters beforehand. But k-means wouldn't be able to determine a prototypical representation for a cluster, as e. g., the k-medoids algorithm would determine. To allow for a more cognitively plausible approach and future applications, a clustering algorithm was selected, which on its own determines the number of clusters.

The *affinity propagation* algorithm was selected for its independence on a predefined number of clusters, while it will determine a so-called *exemplar* representation for each cluster, similar to a median representation. This exemplar representation will be used as a cluster-representing prototype $\tilde{c}^d$. These prototypes can for example be used for speeding up the process of determining cluster affinity for temporary sequences. A new sequence representation may become a member (or exemplar) of a cluster, or trigger the creation of a new cluster representation of level C. To make this clear: clusters are not created on demand, but the affinity propagation algorithm will determine a completely new set of clusters.

Affinity propagation clusters data by identifying subsets of cluster-representative exemplars (Frey and Dueck, 2007). Exemplars are not chosen randomly in order to avoid running into local minima, but are determined by a process of message passing on a similarity matrix of data points, i. e., the action sequence representations.

Two kinds of messages are balanced and weighted to choose exemplars from data points, *responsibility* and *availability* messages. During each iteration of message passing between data points, first responsibilities $r(i,k)$ (see eq. 4.35) are updated using the responsibility message, being sent from data point i to point k, determining how well point k could serve as an exemplar for point i, taking their similarity $s(i,k)$ (see eq. 4.36) into account, which is based on the sequence-comparison method dtw (as discussed in par. 4.2.3). Then availability messages $a(i,k)$ (see eq. 4.35) are collected to determine how fitting it would be for data point i to choose point k as its exemplar. The third step combines availabilities and reponsibilities to determine for data point i, which point k maximizes responsibility and availability $a(i,k) + r(i,k)$, while identifying exemplars if point i and k are the same. The algo-

rithm terminates after a set amount of unchanging iterations. As adapted from Frey and Dueck (2007, pp. 972):

$$a(i,k) \leftarrow \min\left\{0, \sum_{i' \text{s.t.} i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\} \tag{4.34}$$

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{s.t.} k' \neq k} \{a(i,k') + s(i,k')\} \tag{4.35}$$

$$s(i,k) = \text{dtw}(i,k) \tag{4.36}$$

They also discuss that affinity propagation can be viewed as a method that searches for minima of an energy function of a configuration of exemplars. So for application in HPBU, apart from being a quick and exemplar-identifying process, affinity propagation is especially suitable, because the rules for updating exemplars corresponds to minimizing a free energy approximation. This puts the algorithm in the same realm of information-theoretically driven optimization as the rest of the free-energy minimizing hierarchy.

Taken together, the discussed sequence learning and clustering approaches allow the model to make judgements about the suitability of a novel sequence to become represented.

4.3 SUMMARY

This chapter described the modeling of a sensorimotor system on the basis of predictive processing and active inference. We discussed the different functional levels of the hierarchy, how they exchange information with the environment, and with each other. Also, the kinds of actions were described that the model has to cope with during action and perception: handwritten digits.

A very integral part of the chapter took the description of the role of active inference as a possible solution to the problem of motor coordination. Finally, we discussed the self-supervised learning approach for action representations in the form of movement sequences and their hidden variables in the form of clusters. This aspect of the presented modeling approach especially sets it apart from standard hierarchical Bayesian modeling approaches. The minimization of free energy during the different tasks of action and perception was evaluated, which will be discussed later in ch. 6, sec. 6.2. Also, the self-supervised learning approach was evaluated in its recognition performance in sec. 6.1.

5

EXTENDING HPBU WITH A MODEL OF MENTALIZING

The following section describes a predictive-processing based model of a mentalizing system. We will first introduce the mentalizing hierarchy, describing the form of representations and kind of abstractions. This is the part of the complete HPBU model hierarchy that concerns itself with perceiving and reacting to other agents during a situation of collaborative social interaction.

The levels of the hierarchy will be discussed, as well as how they exchange information with their respective next higher and next lower levels. Also, we will cover a description about how a self-other differentiation, as discussed above, based on a sensorimotor sense of agency, can be modeled computationally. We will also come back to how HPBU communicates efficiently (as first discussed in ch. 2), through sensorimotor communication, and the integration of prior experience. Also a part of this section are solutions to specific problems of the presented modeling approach to mentalizing.

5.1 ADDITIONAL MODELING ASSUMPTIONS

In order to model the mentalizing part of HPBU, the set of assumptions that guide computational cognitive modeling need to be extended:

- Predictive processing is a vital step towards conciliating direct social perception with sandwich-model approaches, like theory theory and simulation theory. In that, the hierarchical nature of processing in the brain does not entail a black-box cognitive process. Rather, at work is only the *inferential resonance* of possible *perceptions* with the input. It is able to trigger prediction-error correcting behavior, either to the end of correcting predictions, or to actively change the environment to conform to the expected perception.

- Further, using information of the own body and about its effects on the environment, be it spatially or temporally, is crucial for a sense of agency that further allows to differentiate actions of the self from that of the other.

- We follow the evidence for the link between mentalizing and episodic memory to interpret social interactions in the form of event structures, which we call *coordination sequences*. Coordination sequences can be seen as schemas that contain segments

consisting of mental state belief-attributions. These make it possible to track the belief dynamics between interaction partners during belief coordination over time, up until the interaction goal – a final mental state that is to be reached. They are embedded in the hierarchy on top of the sensorimotor part of the model, so that interaction goals and action schemas that are plausible in the given situation can be predicted.

- By following coordination sequences to their interaction goal, which in effect means that a social interaction of belief coordination can be successfully predicted, prediction error minimizes with every predicted step of the interaction. This can be thought to the extent that this back and forth can lead to a synchrony between the brain states (in the form of probable hypotheses) of interaction partners.

5.2 MODELING A MENTALIZING SYSTEM

Given the modeling assumptions, now a detailed description of the mentalizing part of HPBU will follow, which corresponds to the MENT. Fig. 5.1 depicts an overview over the full model hierarchy.

On top, the mentalizing part includes the G (Goals) and CS (Coordination Sequence) levels. Level G represents clusters of coordinate sequences with similar state-goal pairs of mental states. Level CS represents the sequences of belief-coordinating intentions that connect the state-goal pairs of mental states that are clustered at level G. That is, coordination sequences define a sequence of intentions that lead from a *start* mental state to a *goal* mental state. The intentions that can be triggered in coordination sequences can *reconfigure* the sensorimotor part of HPBU to either produce or observe action. Always, prior beliefs from the Person Model (PM) can influence these processes (see par. 5.2.3).

The kind of intentions that are assumed as being part of a coordination sequence here are threefold:

- First, *intentions to act* trigger the underlying sensorimotor system to produce a belief from PM. PM stores beliefs either in the form of a me-belief, you-belief, or we-belief, and they have their source in the representations of schema level C. So, an intention to act first establishes an intended schema c_I for production, which then automatically triggers subsequent lower-level productions, like s_I at level S (for more detail see par. 5.2.7).

- Second, *intentions to perceive* trigger level CS to wait for a *stable observation* to be communicated from level C. Observation stability here is also defined from an information-theoretic perspective, as the free energy of level C with regard to the currently perceived action sequence.

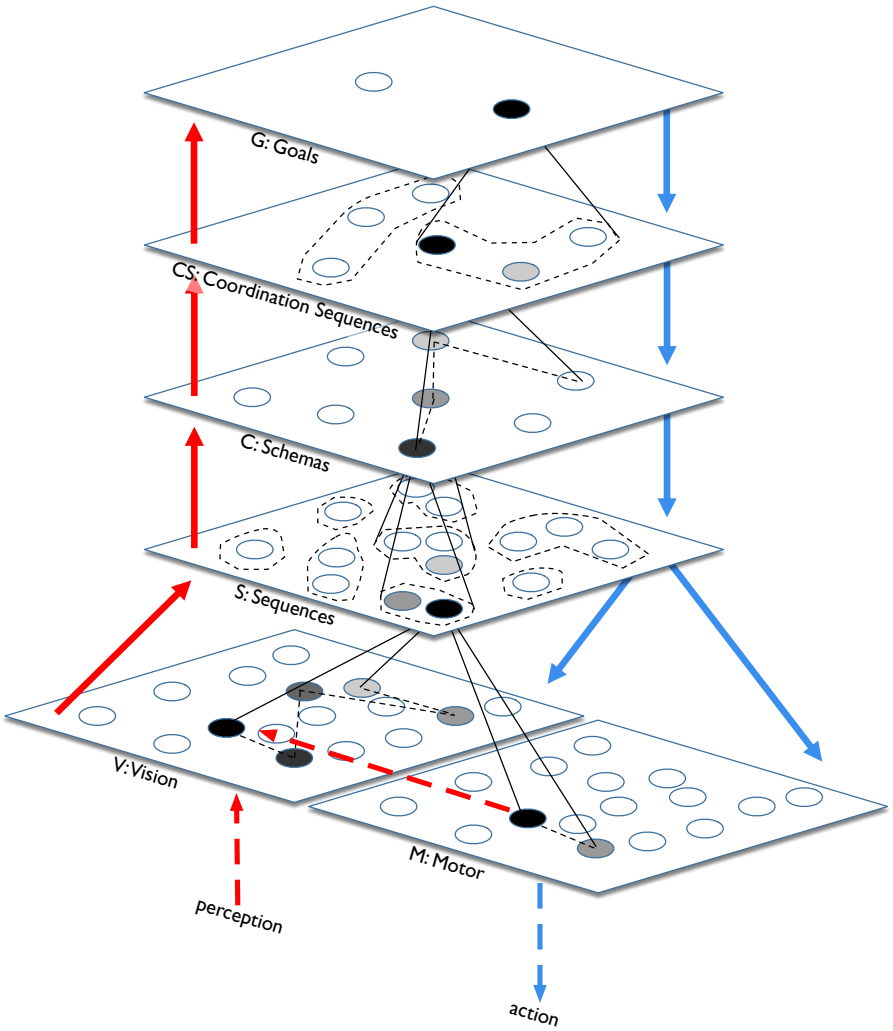


Figure 5.1: Depicted here is the full sensorimotor and mentalizing hierarchy of HPBU. The mentalizing part includes the Goals (G) and Coordination Sequence (CS) levels. Intentions triggered from coordination sequences reconfigure the sensorimotor part of HPBU to either produce or observe action, given certain prior beliefs.

- Third and lastly, *intentions to compare mental states* trigger an internal CS-level comparison of temporarily stored me-beliefs and you-beliefs in PM, with respect to available coordination sequences. Such a comparison is used to decide on whether an interaction goal is reached. In this case, the we-beliefs of the interaction partner are updated to contain the newly established belief. Alternatively, the interaction goal was not reached when it should have been, and a more complex coordination sequence should become available. This in effect prioritizes shorter coordination sequences over longer ones. These can contain repair sequences and intentions to act that allow for sensorimotor communication, which will be discussed later (see par. 5.3.2).

5.2.1 *Extended generative model*

The HPBU hierarchy needs to be extended by a mentalizing part. As already discussed in par. 4.1.3, each level of HPBU's hierarchical generative model maps its internal state space L_i onto the domain of its next lower level L_e. Also, the model maps from external to internal states (L_i) to minimize entropy.

The model described so far will be extended by the mentalizing part. It consists of two levels, the CS level and the G level.

The overall generative model is extended to $(\mathbf{G}, \mathbf{CS}, C, S, M, V)$ (with the sensorimotor part being covered in par. 4.1.3). As described before, the levels are sequentially updated, from its next higher and next lower levels, starting at the top of the hierarchy. Each level then learns to represent and produce the states at the next lower level. Each level of the generative model contains its own discrete probability distribution, where each probability describes a hypothesis of one discrete representation. The joint distribution of the whole generative model hierarchy is similar to,

$$\begin{aligned} &P(\mathcal{G}, \mathcal{CS}, \mathcal{C}, \mathcal{S}, \mathcal{M}, \mathcal{V}, x) = \\ &\qquad P(\mathcal{G}) \cdot P(\mathcal{CS}|\mathcal{G}) \cdot P(\mathcal{C}|\mathcal{CS}) \cdot P(\mathcal{S}|\mathcal{C}) \cdot P(\mathcal{V}|\mathcal{S}) \cdot P(\mathcal{M}|\mathcal{S}) \cdot P(x|\mathcal{M}) \end{aligned} \tag{5.1}$$

5.2.2 *Level definitions and updates*

Fig. 5.2 shows the full HPBU hierarchy, with the highlighted mentalizing part. Predictions from the mentalizing part are sent top-down and compared with (bottom-up) evidence from the sensorimotor part, to drive belief coordination updates within the hierarchy.

The G level contains clusters of coordination sequences with similar state-goal pairs. The CS level below that provides event structures for belief coordination, leading towards the goal state.

As described in detail (see sec. 3.3.1), the mentalizing structures proposed in this work, functionally overlap with event structures. That is, imaging studies show reliable activations of the mentalizing region mPFC during retrieval from episodic memory (Hassabis and Maguire, 2007; Maguire and Mummery, 1999). Also, meta-analyses on episodic simulation report an overlap between brain regions associated with episodic memory, and the default mode network (Benoit and Schacter, 2015). Specifically, area TPJ has been associated with inferring intentions in social situations (Van Overwalle, 2009).

5.2.3 *Person model and its influence*

In fig. 5.2, seemingly at the top (or lower-left) of this hierarchy you find the PM, which here should not be confused with being a level of the hierarchy. Rather, it takes the function of a partially permanent

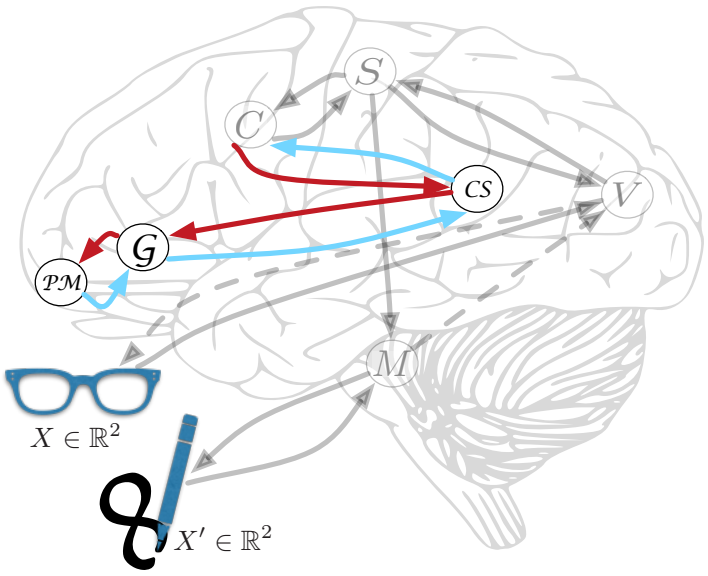


Figure 5.2: Depicted is the mentalizing part of HPBU. Predictions are sent top-down and compared with sensory (bottom-up) evidence, to drive belief updates. All levels have loose associations with the displayed cortical structures. In the person model (PM) differentiated mental-state attributions are stored. The goals (G) level contains clusters of coordination sequences with similar state-goal pairs. The coordination sequence (CS) level below that provides event structures for belief coordination, leading towards the goal state. The mental states to which these structures are compared to are updated from the sensorimotor part of HPBU.

and partially temporary storage of beliefs. For one, PM stores prior information of specific interaction partners, i. e., about the already established common ground with that interaction partner (see par. 3.3.2 for the background). Also, PM stores temporary beliefs of the ongoing interaction, holding the belief of the agent itself (called *me-belief*), and beliefs about the underlying action intention of the interaction partner (the *you-belief*). These beliefs are communicated to PM from current stable beliefs at level C. It also stores the goal belief that the interaction strives for (from the agent's perspective), or in other words, the communication goal (here called *we-belief*).

Two aspects are important to understand these different kinds of beliefs stored in PM: one is the interaction's goal-state, which will also be represented as a we-belief. All interaction partners should strive to achieve this we-belief, but not every agent is aware of it from the beginning of the interaction. As we will later discuss in more detail for our evaluation, agents will take on different roles in a communication game, either that of a *leader* or a *follower*. Only the agent with the leader role defines the we-belief from the beginning, and strives for that belief to be established as common ground. The other aspect

is the belief that is attributed to another agent – as the you-belief – inferred from perceived behavior. If that belief is similar enough to the we-belief, then belief coordination has been successful.

PM is set up in such a way that it influences levels of the hierarchy using long-range connections. As such, prior information (in the form of established common ground) about an interaction partner, influences the belief update of level S. The discrete probability distribution in S represents the probabily of perceiving or producing an action sequence. Thus, PM can increase the chance of perceiving an action for which information was already shared between interaction partners, i. e., perception can be biased toward prior information using already established we-beliefs.

Prior information pm about an interaction partner q is stored as $pm_q \in PM$. Each pm_q contains N_q action sequences that are in common ground with that specific interaction partner, with $N_q = |pm_q|$. Thus, each pm_q here contains only the *we-belief* of prior interactions. To maintain compatibility with the probability distribution of all action sequences $\mathcal{S}$ at level S, the probability distribution P_{pm_q} will be constructed for all s_m using $f_{pm} : \mathbb{R} \to \mathbb{R}$:

$$P_{pm_q}(s_m) = f_{pm}(s_m) \quad \forall s_m \in S \tag{5.2}$$

$$f_{pm}(s_m) = \begin{cases} \frac{1}{N_q} & \text{if } s_m \in pm_q \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

P_{pm_q} will then influence top-down beliefs P_{td} prior to the belief update (see eq. 4.6) at level S,

$$P'_{td}(\mathcal{S}) = P_{td}(\mathcal{S}) + K_q(P_{pm_q} - P_{td}(\mathcal{S})), \tag{5.4}$$

with a partner-specific Kalman gain K_q that models the trust in the interaction partner. K_q can take many values, but here it is equally weighting the influence of each interaction partner, for which the established we-beliefs are stored in PM.

Thus, upon a successful belief coordination, the established we-belief is stored in PM, for it to influence level S in future interactions. The me-, and you-beliefs, important for tracking an ongoing belief coordination, are not influenced by the described process of influencing level S beliefs.

5.2.4 *Levels and representations*

State-goal pair: the mental states at the beginning and end of a coordination sequence.

As you can see in fig. 5.3 at the top of the mentalizing hierarchy, we find the G level. It represents abstract clusters of similar coordination sequences $\{g_1, \ldots, g_p\}$ that contain similar state-goal pairs. *State-goal*

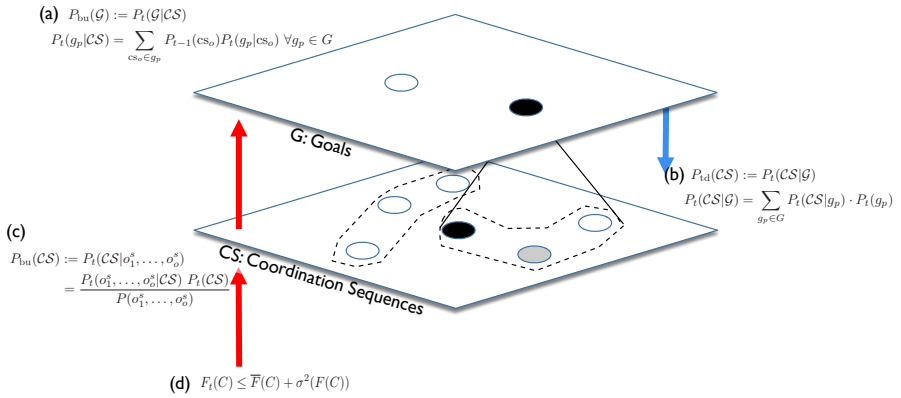


Figure 5.3: This technical overview shows the following input and output calculations embedded at levels CS and G: (a) describes the bottom-up update of $\mathcal{G}$ (approximately similar to eq. 4.13), while (b) describes the top-down update of $\mathcal{CS}$ (similar to eq. 4.15), and (c) describes its bottom-up update (approximately similar to eq. 4.16). (d) describes the detection of stable observations (par. 5.2.5).

pairs represent the mental states at the beginning and end of a coordination sequence, clustering coordination sequences with the same interaction outcome (see sec. 3.3.1). G contains a discrete probability distribution $\mathcal{G}$ over p discrete states, and maps to it's next lower level *CS*, with $cs : G \mapsto CS$. Each G-level represenation g_i clusters coordination sequences $CS' \subseteq CS$ by similar state-goal pairs. Similarity here depends on similar start-, and goal-state mental states of the corresponding coordination sequences. The goal state is a final mental state in a coordination sequence, which has to be reached, for a coordination sequence to be successful.

Level G calculates its bottom-up posterior using the same soft evidence method, as was described for level C (see eq. 4.13). Here, since G represents the top of the hierarchy we have no top-down posterior, so that the level G posterior belief update will be calculated with the posterior from the last time step $P_{t-1}(\mathcal{G})$, in place of the top-down posterior $P_{td}(\mathcal{G})$. The same is true for calculating free energy.

Below that, the CS level represents so-called coordination sequences, $\{cs_1, \ldots, cs_o\}$, which provide event structures that were described, along with their neuroanatomic associations, in sec. 3.3.1. Shortly summarized, it defines the progression of a belief-coordinating interaction that allows to move from one set of mental states to another (e. g., the respective start-, and gloal-states, or state-goal pairs of G). Each coordination sequence cs_o contains a sequence of intentions that each may also allow to entail a strategic biasing of precision weighting (as we will discuss soon in par. 5.2.7). CS contains a discrete probability distribution $\mathcal{CS}$ over o discrete states, and maps to it's next lower level C, with $c : CS \mapsto C$.

Level CS calculates its top-down posterior similarly to level S, from a mixture of experts (see eq. 4.15).

5.2.5 *Comparing coordination sequences*

From level C only *stable* observations (o^s) will be communicated to CS, such that free energy at level C is:

$$F_t(C) \leqslant \overline{F}(C) + \sigma^2(F(C)) \tag{5.5}$$

For the bottom-up posterior only these stable observations in the form of one maximum posterior hypothesis from level C are communicated to PM, to be stored either as a me-belief or a you-belief, depending on the amount of SoA attributed to it (see eq. 5.9).

Similarly to the sequence comparison at level S, the coordination sequence likelihood $P(cs'|cs_o)$ is the coordination sequence difference dtw_{cs} between known coordination sequences and a temporary coordination sequence cs' that grows over the interaction time. The sequence difference is same as dtw (see eq. 4.28), with the minor change that $d^{cs}_{(a_i,b_j)}$ adds zero only if there are no differences between mental states *i* of coordination sequence *a* compared with mental states *j* of coordination sequence *b*:

$$d^{cs}_{(a_i,b_j)} = \begin{cases} 0 & \text{if } a_i^{me} = b_j^{me} \wedge\ a_i^{you} = b_j^{you} \\ 1 & otherwise. \end{cases} \tag{5.6}$$

CS and G-level representations are not learned by the implementation of the model presented here. Instead, a number of coordination sequences have been defined by hand, in order to allow for a number of increasingly complex interactions. These include sensorimotor communication strategies and the inclusion of prior beliefs about another agent, for adaptive reciprocity based on previously perceived false beliefs. These are then collected into different Goal level cluster representations that code for similar state-goal pairs. The coordination sequences that were used will later be specified in detail for the evaluation simulations (see fig. 6.15).

5.2.6 *Meta-communication*

HPBU also has so-called meta-communicative acts at its disposal, i. e., they can also be perceived and produced as intentional acts. These are not represented at the schema level. Rather, one would probably represent them in other hierarchical representations that equally allow for their perception and production. For this work, meta-communicative acts are represented as stand-in (or virtual) hierarchies for production and perception, i. e., they are implemented purely as

data signals to be perceived with certainty, and exchanged between the implemented HPBU agents.

Meta-communicative acts, represented here, are the following:

- *Social gaze* is shared between interaction partners to establish a direct social interaction once at the beginning of the interaction. This allows to instantiate mutual gaze between interaction partners. We discussed its importance for social interaction, motor contagion, and the general feeling of direct social interaction, in par. 2.2.3.

- *Thumbs-up* is a signal to be communicated to end communication for good. In human communication, other signals are often used, e. g., nodding, a salient smile, or specific other gestures. Here, we choose the salient thumbs-up iconic gesture to signal that the signaling agent thinks that the interaction could end here. All coordination sequences end with this meta-communicative signal.

Now, with meta-communicative acts in place, coordination sequences are able to trigger intentions to observe or to act toward reaching their goal state. The next question is: how does the sensorimotor part of HPBU react to such signals that carry the intention to observe or to act?

5.2.7 *Intentions to act and intentions to observe*

For the sensorimotor part of HPBU, intentions to observe and intentions to act do not represent automatic responses that are naturally evoked from sensorimotor representations. Rather, they are strategically placed biases that come from outside the sensorimotor part: the mentalizing hierarchy.

The sensorimotor system responds to these biases (as introduced earlier as bias *b* in par. 4.2.4) and reconfigures its driving signal (focusing on perception or action), in order to allow the minimization of free energy in the mentalizing part of HPBU. Without this reconfiguration level CS and G would not or only by chance be able to reduce prediction error.

The *intention to act*, strategically placed at level CS, for one sends a strongly biased distribution as a *prediction* to level C, where it is used to update its top-down posterior. In a way this is similar to *clamping*, as it is used in neural network architectures, where a neuron has its value forcibly fixed to a certain value. Such a clamped neuron is often used as an input unit for the network, e. g., to force a generative model to produce a certain behavior. Also, the intention to act places an intention for the to-be produced belief in the long-range connection to levels C and S, thus triggering the production of the intended action in the sensorimotor part of HPBU. The intention stems either from the me-belief or the you-belief in PM, depending on whether the agent's

own belief is to be produced (me-belief), or the other's observed belief (you-belief) is to be reciprocated. In addition, the intention to act helps the sensorimotor part to achieve its goal to minimize free energy, given the new constraints, by biasing the Kalman gain K (see par. 4.2.4) with a low b for the level posterior belief update towards maintaining the prior, i. e., the top-down posterior.

The *intention to observe*, in turn, also strategically sends a prediction of a distribution of level C representations. There, the probabilities are clamped to the one belief from PM that is to be observed (me-belief or you-belief). In addition, the intention to observe helps the sensorimotor part to achieve its goal to minimize free energy, given the new constraints, by biasing the Kalman gain K with a high b for the level posterior belief update towards the observation, i. e., the bottom-up posterior.

Now that the full HPBU hierarchy is in place, how does it allow for belief coordination that is not only achieved by a reciprocal back and forth, but that is also efficient?

5.3 EFFICIENT BELIEF COORDINATION

Based on free energy minimization HPBU allows to engage in a process of belief coordination in a reciprocal back and forth of the inferred intentions underlying perceived behavior. This is done in a process of inferential resonance of possibly perceived behavior that is also in a second step interpreted in a process of inferential resonance of possible communicative goals. Both can in response trigger the engagement with an interaction partner based on inferred intentions, by putting the pressure to minimize free energy on the direct interaction with the environment.

Both, resonance and reciprocity, allow for a reciprocal back and forth, although by itself this is not an example of efficient communication. For that, the incorporation of inferred beliefs in future decision making and model selection is necessary. Also, prior information about the interaction partner should influence perception.

These are strategies for efficient communication that may lead to an update of common ground. But neither would be possible without a reliable differentiation of perceived behavior to be the consequence of action, either from the self or another agent. This information is used to attribute actions, as inferred at level C, to either beliefs of *self* or beliefs of *other* in the person model.

5.3.1 *A model of sensorimotor sense of agency*

During online social interaction, the sensorimotor system potentially gets involved in simultaneous action perception and production processes. This makes the correct attribution of agency to perceived action effects necessary. As suggested in the background on SoA (see sec. 3.4), there are two kinds of processes that lead to informative cues: a predictive and a postdictive process. Here, we models these accounts and integrate them into a sensorimotor SoA for produced actions, which will depend on the likelihood calculated at the sequence level S: in the predictive process, we calculate the likelihood of the perceived action sequence s', given the predicted action sequence $s_m \in S$ in the sequence comparison function $\mathrm{dtw}(s', s_m)$. In the postdictive process, we have the intention to act and the delay in the action-outcome for temporal integration. This temporal integration depends on the predicted and perceived temporal delay of the predicted action, and the sequence level's precision. Precision in this context will stretch or sharpen the likelihood of temporal integration (see eq. 4.30).

Following the evidence for a fluency effect that accumulates the repeated success in correctly predicting and selecting actions (Chambon et al., 2014), this accumulation of evidence is also modeled. That is, the likelihood of the current action of the intended sequence s_I is put into a Kalman filter to estimate the agency (see eq. 5.9). To make this clear, SoA is calculated for the given intended action sequence at level S. The Kalman filter estimates the agency $\hat{a}_t$ from the likelihood $P(s'|s_I)$ and the previous agency estimate $\hat{a}_{t-1}$. If no intended sequence is available ($P = 0$) the agency will slowly decrease. Kalman gain K_t is calculated from the sequence level's free energy F_S and precision π_S.

$$\hat{a}_t = \hat{a}_{t-1} + K_t(S)\ (P(s'|s_I) - \hat{a}_{t-1}) \tag{5.7}$$

$$P(s'|s_I) = \mathrm{dtw}(s', s_I) \cdot e^{-\frac{(\Delta t_{s'} - \Delta t_{s_I})^2}{2\ \pi_S^2}} \tag{5.8}$$

$$K_t(S) = \frac{F_S}{F_S + \pi_S} \tag{5.9}$$

By allowing the agency estimate only to accumulate through this filter, strong fluctuations are dampened. Further, with the gain governed by precision and free energy, the influence of the estimate will strongly depend on the success of previous predictions.

The essential elements for the sense of agency of the perception-action loop are an intention signal for a specific action production, the correct prediction of the learned action, and its timing. At level S of HPBU, the prediction and evaluation of an action and its timing are embedded in each sequence. The intention signal for a specific action production can essentially be described as a high precision predictive

This subsection on the mechanisms of sense of agency for self-other distinction was adapted from a section in a previous publication in Kahl and Kopp (2018).

corolary discharge (see par. 4.2.4) that is very strong and stable over time.

If it is then the case that such a high precision prediction is the driving signal, and the probability for the predicted sequence stays low, the model's free energy will be high. An interpretation of a high SoA is that either something unpredicted is influencing the action production, or it is not the system's production at all that is perceived. If so, the perceived action-outcome can be attributed to another agent.

With a model of SoA and self-other differentiation available, future behavior can be guided by inferred beliefs about the other, e. g., in sensorimotor communication.

5.3.2 *Sensorimotor communication*

Sensorimotor communication, as previously modeled, is an optimization problem over an action trajectory to be as different – or informative – as possible without losing the vital information necessary to be classified as the original trajectory's interpretation (Pezzulo et al., 2013).

In HPBU sensorimotor communication also has the purpose to select or alter an action representation in order to tailor the next communicative act to the interaction partner's needs. These needs have to be met by selecting an action sequence, and are defined, e. g., by prior information, representations that differ in terms of sequence representations, or also schema membership.

We will soon come to the evaluation chapter, where instances of the model will be put into interacting agents. In a scenario that allows for belief coordination to happen, agents will take on different roles in a communication game, either that of a *leader* or a *follower*. With these roles come different kinds of knowledge, specifically, only the leader agent will have a me-belief from the start, along with the intention to communicate that belief to other agents, while taking care that they have understood correctly.

The sensorimotor communication strategy applied here, for selecting an action sequence representation for a next communicative act, is based on two aspects:

One is the interaction's goal state, which will also be represented as a we-belief, and can be understood as a goal belief. All interaction partners should strive to achieve this we-belief, but not every agent is aware of it from the beginning of the interaction. Only the agent with the leader role defines the we-belief from the beginning, and strives for that belief to be established as common ground. The other aspect concerns the belief that is attributed to another agent – as the you-belief – inferred from perceived behavior, which is being produced by

interaction partners. If the you-belief is similar enough to the we-belief of the goal state, then belief coordination has been successful.

In cases of unsuccessful belief coordination, a strategy of repair has to be implemented. In this strategy, the schema of the other's you-belief is understood to be a *distractor* (c^d), i. e., it distracts the other's perceptual inference from the correct inference (which is the one communicated by the leader agent). As the next step, a new action sequence from the goal belief's schema cluster has to be chosen. Here, a mode of action-sequence selection is applied that is different than the random selection, which normally selects a member of an intended cluster for production. This time, the action sequence is chosen conditionally on being *most different* from the distractor schema.

Thus, the subset of action sequences clustered under the goal belief's (c^g) schema $s_m \in c^g$ is compared with the distractor's prototype $\tilde{c}^d$, in order to find the most different sequence that becomes the new intended action sequence s_I:

$$s_I = \underset{s_m \in c^g}{\arg\max}\, \mathrm{dtw}(s_m, \tilde{c}^d) \tag{5.10}$$

With a new s_I found, it becomes the new sequence to be produced, and hence communicates the goal belief c^g to the interaction partner in a way that is strategically tailored to take its prior beliefs and perceptual differences into account. Of course, this repair strategy cannot account for the possibility of a missing representation in the interaction partner. A missing representation would make the intended understanding impossible, unless learning were enabled during such situations in social interaction.

5.4 SUMMARY

Now that the full model of hierarchical predictive belief update is in place, let us summarize it. To do so, we briefly revisit chapter 2, but with the presented model in mind. There I proposed that we as humans face three questions before and during social interaction:

First, in order to engage in social interaction we have to identify an interaction partner as such, or identify ourselves as being the receiver of communication. Here, HPBU provides the possibility to enage in direct social interaction by addressing an interaction partner using social gaze as a meta-communicative signal. An interaction for belief coordination can be established once the addressed agent returns the social gaze, so that mutual gaze is established. This allows for a – possibly very direct – social perception of communicative acts.

The second question was about getting a grasp on the mental states of our interaction partner, as this is the goal of the process of updating common ground. HPBU allows to attain this grasp, including a differentiation between the self and an interaction partner's behavior, with a process of inferential resonance of possible perceptions

with the perceived behavior. This is able to trigger prediction-error reducing behavior, either to the end of correcting predictions, or to actively engage with the interaction partner's beliefs through action production, i. e., active inference. Through this hierarchical process of resonating explanations for behavior, and resonating explanations for the interaction goal, common ground can be established.

The third question was concerned with answering when an interaction can be deemed to have ended, successfully or not. This process of reciprocal belief coordination seems imperfect, as it has been repeatedly shown that intial interpretations can linger, supported by contextual information, only being good-enough for the interaction to go on. But in general, misunderstandings and communication errors are not fatal for social interactions, but can be repaired. HPBU supports this process by allowing for a strategy of efficient belief coordination repairs, i. e., the support for sensorimotor communication, as a means to engage with an inferred false belief in the interaction partner. For that, the interaction goal will again be produced, but with strategically altered action kinematics to allow for better differentiation from the false belief. This makes belief coordination and error correction possible, so that a good enough common ground can be established between interaction partners.

Of course, we still do not have a full understanding of all details of how social cognition is implemented in the brain, but HPBU may be a viable framework for belief coordination, and a possible example for an algorithmic explanation based on free energy minimization. In the following, several evaluation attempts will be described in order to establish HPBU's viability.

6 RESULTS

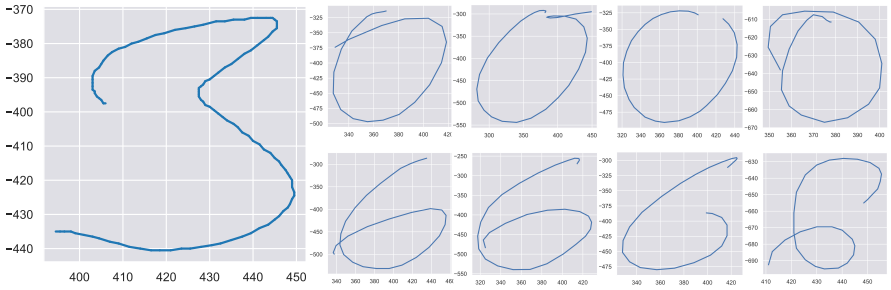
In the previous chapters we reviewed the theoretical background for human social interaction and belief coordination from the perspectives of linguistics, conversation analysis, cognitive neuroscience, and cognitive psychology. Also, we presented a computational modeling approach to sensorimotor and mentalizing processes called HPBU in detail.

In the following, we will desribe the results of several evaluations of different aspects of the HPBU model. For now, the focus will be only on the sensorimotor part of HPBU. First of all, the self-supervised learning approach needs to be evaluated with regard to its recognition performance (sec. 6.1). Then, the model's ability to minimize free energy is evaluted during perception of action sequences as well as during action production (sec. 6.2). After that, the focus will move towards a social perspective, while staying put on the sensorimotor part. There, the correct differentiation of self from other will be evaluated, and by that the model of sense of agency, underlying the self-attribution judgement (sec. 6.3). Finally, we will focus on the whole HPBU model, integrating the sensorimotor part with the mentalizing part. To evaluate the mentalizing model in interaction with the sensorimotor part, its capabilities will be tested in a setting of belief coordination, in interaction with multiple other agents (sec. 6.4).

6.1 MODEL RECOGNITION PERFORMANCE

Here, we evaluate the recognition performance of the self-supervised approach implemented in the sensorimotor part of HPBU.

The data the model trains on is the handwriting corpus, described in sec. 4.1.2, which contains the recordings of handwritten digits (*0-9*) from five user sources. The corpus can be split into sets of user-specific handwritings from the five different user sources, which we call *a*, *b*, *c*, *d*, and *e*. We will here not evaluate trainings or tests on user source *e*, as this user's writings were performed at too slow speeds, such that resampling, smoothing and rescaling would have been needed. For an example, see fig. 6.1 *(a)*, showing the writing of a 3. No resampling or rescaling was necessary for data of user sources *a-d*. Nevertheless, certain digits were not uniformly written by people, making their reliable classification unlikely, and reducing the overall achievable accuracy. For examples, have a look at fig. 6.1 *(b)*, showing writing examples of the digits 0 and 6. Thus, I would expect to see the overall achievable accuracy being reduced by these two classes having a high



(a) User source *e* distortion example

(b) Corpus variations examples

Figure 6.1: (a) User source *e* example. This user's writings will not be used, as they were performed at too slow speeds that creates visible noise, such that resampling, smoothing and rescaling would have been needed. (b) Certain digits were not uniformly written by people, making their reliable classification unlikely. The digits 0 and 6 are such examples.

risk of being wrongly classified as one or the other. This reduces the overall achievable accuracy for 10 different classes by 0.1-0.2, leaving an expected achievable accuracy of 0.8-0.9.

The available user sources allow to mix the different handwritings to evaluate the recognition performance on similar (same user source) or unseen data (different user sources). The focus here will primarily lie on the performance on unseen data, as this gives us information about the classifier's generalization performance.

We will compare two classification training and evaluation approaches: 1) One is a training on data from one specific user source, up to three epochs (repetitions). Each trained classifier is tested with data from a specific user source, to evaluate its generalization performance. 2) The other is a training on a pooling of all available data from the four different user sources. Here, a classifier is trained once on user sources *a*, *b*, *c*, and *d*. Its generalization is tested on each of the available user sources.

During every training run, HPBU will decide – based on two methods for the detection of a salient segment (see sec. 4.2.5) – whether a new representation should be learned. This way, it is not guaranteed that a single training run will allow the model to get a grasp on (and represent) the full corpus data, which often leaves out several exemplars from the training set untouched. With repeated training, there is of course the threat of overfitting to keep in mind.

Thus, first the effect of repeated training on user source *a* will be shown in fig. 6.2, where you see a confusion matrix. It shows the number of correct classifications on its diagonal, with an increase of correct classifications with repeated training.

Next, we will look at the classification performance on unseen data from different user sources, in order to evaluate the model's

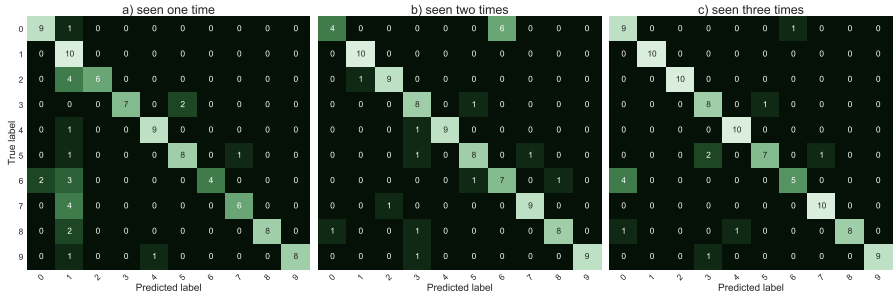


Figure 6.2: The figure shows confusion matrices after repeated training and testing on user source *a*. a) Shows the classification performance after seeing the data once, with more misclassifications as when b) the data had been seen two times, or c) three times, where the number of correct classifications is the highest.

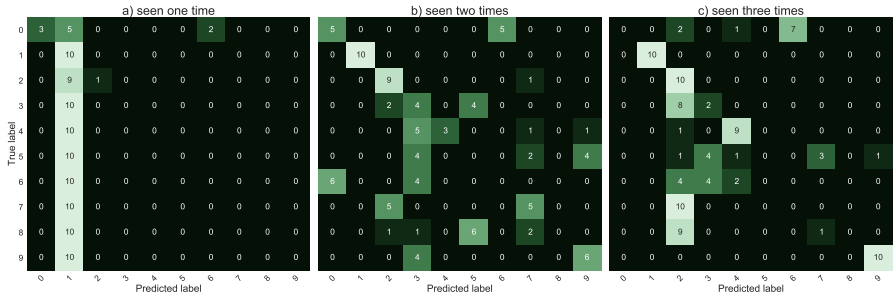


Figure 6.3: The figure shows confusion matrices after repeated training on user source *a* and testing its generalization to user source *b*.

ability to generalize from one user source to others. We will also evaluate the influence repeated training on a single user source has on the generalization performance. In fig. 6.3 confusion matrices are shown for the model repeatedly trained on user source *a*, testing its generalization to user source *b*. Similarly, in fig. 6.4 confusion matrices are shown, the model repeatedly trained on user source *a*, testing its generalization to user source *c*. And finally, in fig. 6.5 confusion matrices are shown, for the repeatedly trained model, tested on user source *d*.

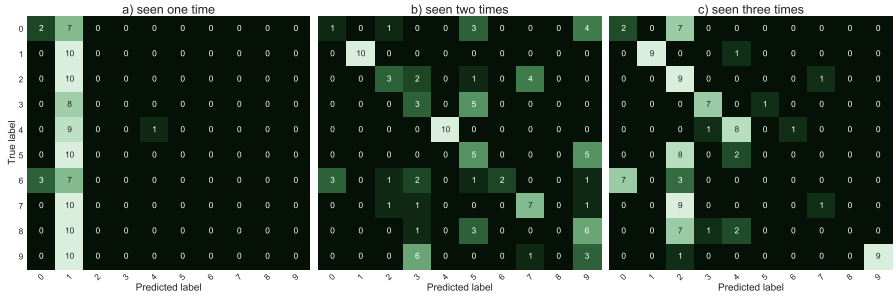


Figure 6.4: The figure shows confusion matrices after repeated training on user source *a* and testing its generalization to user source *c*.

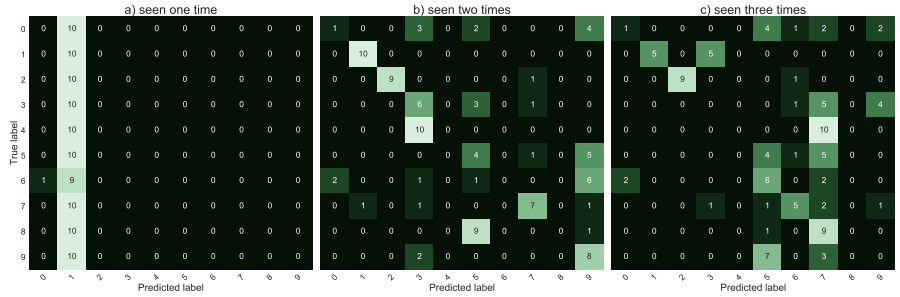


Figure 6.5: The figure shows confusion matrices after repeated training on user source *a* and testing its generalization to user source *d*.

In fig. 6.3, 6.4, 6.5, a general increase in classification performance can be seen, which speaks for the influence of repeated training on generalization, but overall generalization is lacking for all three tests.

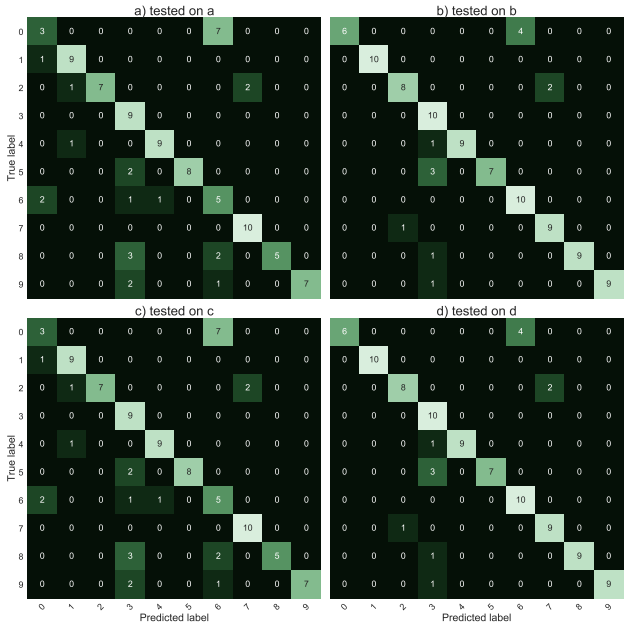


Figure 6.6: The figure shows confusion matrices resulting from training on a pooling of user source (*a*, *b*, *c*, and *d*) and testing its performance to classify each user source.

The second approach to be tested is the pooling of training on different user sources. We have just seen that repeated training allows to increase classification performance, and – to some degree – increases generalization performance. Fig. 6.6 shows the performance of HPBU classification, trained on a pooling of user sources, i. e., trained once on user sources *a*, *b*, *c*, and *d*. The general classification performance of this model, trained on a pooling of user sources, shows that it correctly classifies most of the seen tests.

For better comparison of the classification performance, the micro-average accuracy was calculated for each classifier. This means that the resulting average accuracy aggregates the contributions of all classes

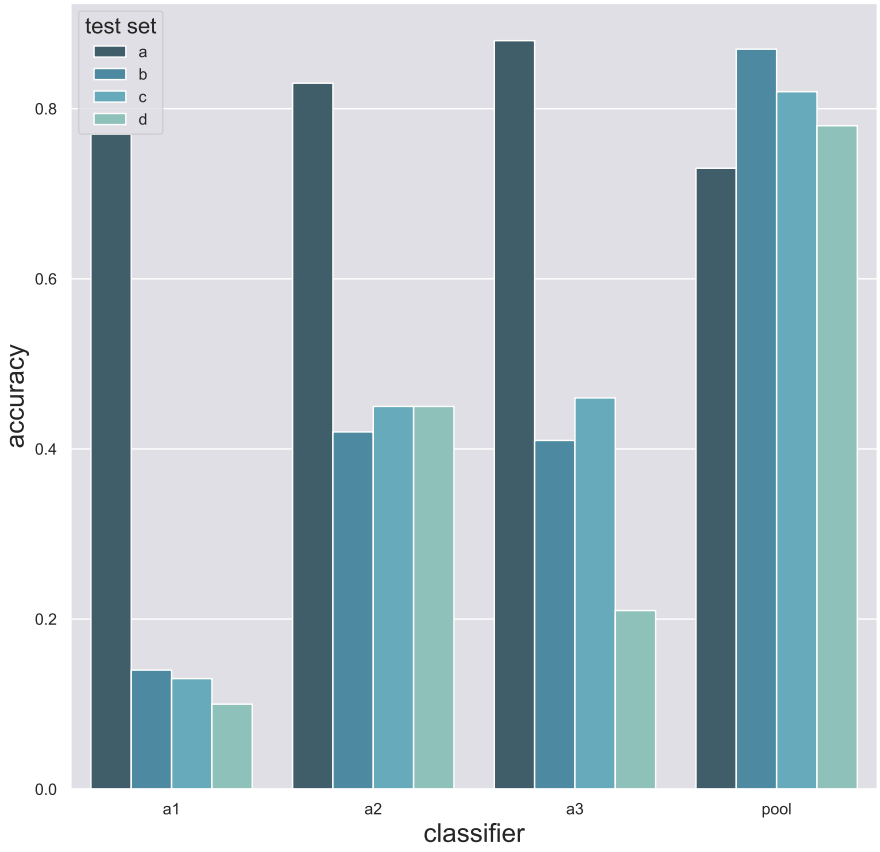


Figure 6.7: The figure shows the micro-average accuracy of each classifier, tested separately on each user source. Each classifier is described using the user source of its training and its number of repeated training runs (e. g., *a1* has been trained on user source *a* and trained only once).

to compute the average metric. The result can be seen in fig. 6.7. There, each classifier is described using the user source of its training and its number of repeated training runs. For example, *a1* has been trained on user source *a* and trained only once, while *a2* has been repeatedly trained twice.

In the comparison of the accuracy of the different classifiers, each tested on the different user sources, the best generalization can be achieved by the *pool* model trained on the pooled set of user sources. The repeated training classifier *a2* seems to generalize the best to other user sources, without a decrease in accuracy, as can be seen from *a3* when tested on user source *d*. In my understanding, the minimal increase of performance from testing on user source *a*, while losing accuracy for testing on user source *d* is a sign of overfitting, which should be avoided. Overall, the pooled training set model seems to perform best.

This comes at a caveat, as you might expect. The different models contain a very different amount of representations, as described in

tab. 6.1. This is especially true for the model trained on the pooled user sources, where 347 of the seen 400 sequences (from four user sources) have become represented in 53 clusters. In the current implementation, without parallelization and optimizations of sequence comparisons, the amount of necessary sequence comparisons is so high that a significant increase of necessary compute power would be needed to still achieve online-performance.

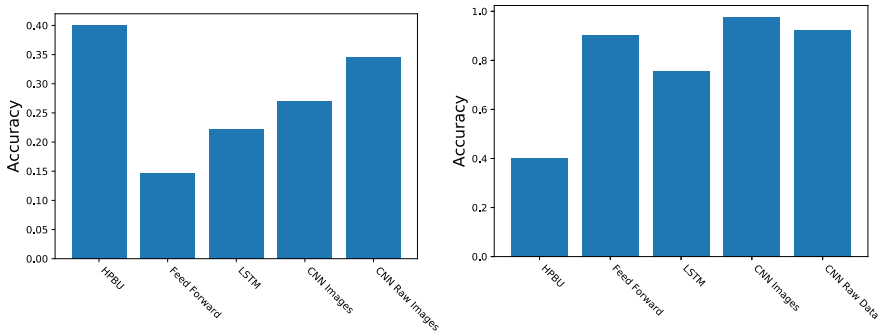
CLASSIFIER	NUMBER OF SEQUENCES	NUMBER OF SCHEMAS
a1	67	13
a2	94	25
a3	105	27
pool	347	53

Table 6.1: Shown are the number of representations in the Sequence and the Schema levels, for each evaluated classifier. The number of representations can lead to an increase of necessary compute power to still achieve online-performance.

COMPARISON WITH OTHER MODELS HPBU was also used in a master's thesis, where its performance on the described corpus was compared with different connectionist models. There, state-of-the-art Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) and LSTM networks were trained on the same corpus of handwritten data, as HPBU. Also, a CNN was created by AutoML to create a gold standard for the described corpus of handwritten digits. The method called AutoML automatically creates and tests different model architectures and parameters (e.g., Real et al., 2017). In this work though, all the user sources were trained upon, even though source *e* has the disadvantage of containing 4 times the number of samples per digit, including a lot of noise due to the slow writing speed. After a single epoch (seeing the whole training set once), HPBU has a lead on the connectionist models in terms of accuracy, which was then not tested any further in this master's thesis due to technical problems with repeated training of HPBU (that have since been resolved). The four connectionist models were trained with a stochastic optimizer (called Adam, see Kingma and Ba, 2017), and either received images of the plotted trajectories or received the trajectories in the form of normalized arrays (please see fig. 6.8*). The connectionist models could then, after 50 epochs, beat the initial lead of HPBU. The AutoML model performs best, with an accuracy of about 0.78 (please see fig. 6.9†). The AutoML generated CNN

* Voß, Hendric (2019): HPBU vs deep learning approaches - first epoch. figshare. Figure. https://doi.org/10.6084/m9.figshare.10252892.v1

† Voß, Hendric (2019): HPBU vs deep learning approaches - AutoML comparison. figshare. Figure. https://doi.org/10.6084/m9.figshare.10028972.v1



(a) After first epoch (b) Final model comparison

Figure 6.8: (a) Accuracy of the five different networks after the first epoch. The four networks on the right were trained with an Adam optimizer. *CNN Images* was trained with plotted trajectories, while *CNN Raw Images* received the information in the trajectories as a normalized array. (b) Comparison of the accuracy after 50 epochs of learning, while HPBU could not be repeatedly trained.

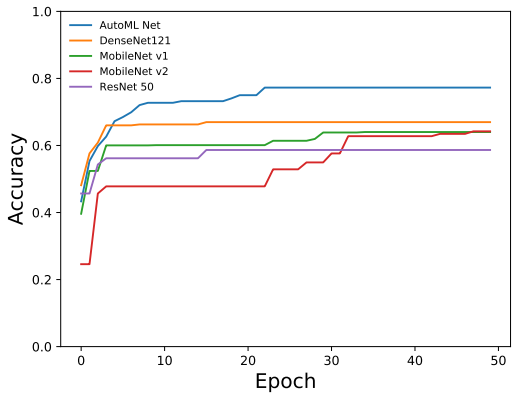


Figure 6.9: The accuracy by epoch of several known machine learning models against one model generated with a custom AutoML algorithm.

model was compared with other known machine learning models on the full corpus of trajectories of handwritten digits. This is very close to the expected maximal accuracy of 0.8-0.9. In the master's thesis, similarly the observation was made that digits 0 and 6 were often confused. Also the digit 5 was identified as an outlier. The discrepancy between classification accuracies of HPBU in the model comparison from the master's thesis (see fig. 6.8 (a)) and the here presented test results (see fig. 6.7) may very well result from leaving out user source *e*.

With the recognition performance established we can now say that at least to some degree the correct recognition of perceived handwriting can be assumed, with the repeatedly trained model *a2* having a very good accuracy on the user source it was trained on. It generalizes

to other user sources without overfitting. *a2* is also a good choice because it still is much more computationally inexpensive than the *pooled* model, trained on all available data.

6.2 FREE ENERGY MINIMIZATION FOR ACTION AND PERCEPTION

With the recognition performance of handwriting established, now the model's general ability to minimize free energy can be evaluated. The ability of the generative model to minimize free energy during perception of action sequences is quite vital to the overall viability of the modeling approach. For example, if the model would fail to infer the correct action sequence or schema, despite minimizing free energy, or would actually infer the correct schema, while free energy could not be minimized, this would point to a serious failure of the model to capture the correct information-theoretic irregularities. Either that, or the general approach to perceptual inference through prediction-error minimization would need to be called into question.

Over time the writing samples are fed into the model as a *visual input stream*. Level V reacts to the visual input stream, with its representation's likelihood changing with the drawing angle of the input drawing. That likelihood updates level V's posterior belief, by also incorporating the predictions from the level above. There, level S also resonates to the salient movements detected at level V, comparing them to known action sequences, while incorporating level C predictions. At level C schemas are evaluated based on their prior probability, and based on the action sequence probabilities of each schema's cluster.

Here, three expectations of the model's ability to minimize free energy are evaluated:

- First, the perception of a *known* handwriting example should minimize free energy.
- Second, the perception of an *unknown* handwriting example should not lead to minimized free energy.
- Third, the production of an action should lead to minimized free energy.

EVALUATION In fig. 6.10 you can see three columns of model dynamics being plotted. Each row shows the posterior probability dynamics, plotted as heat maps over time (color coded from dark green to white, from belief probabilities $P = 0$ to $P = 0.6$ for best differentiability), while the model works to minimize free energy, given its presented task. Each heatmap shows the level's posterior probabilities after beliefs are updated bottom-up and top-down. Also, below the

† The evaluation of free energy minimization was previously discussed in Kahl and Kopp (2018).

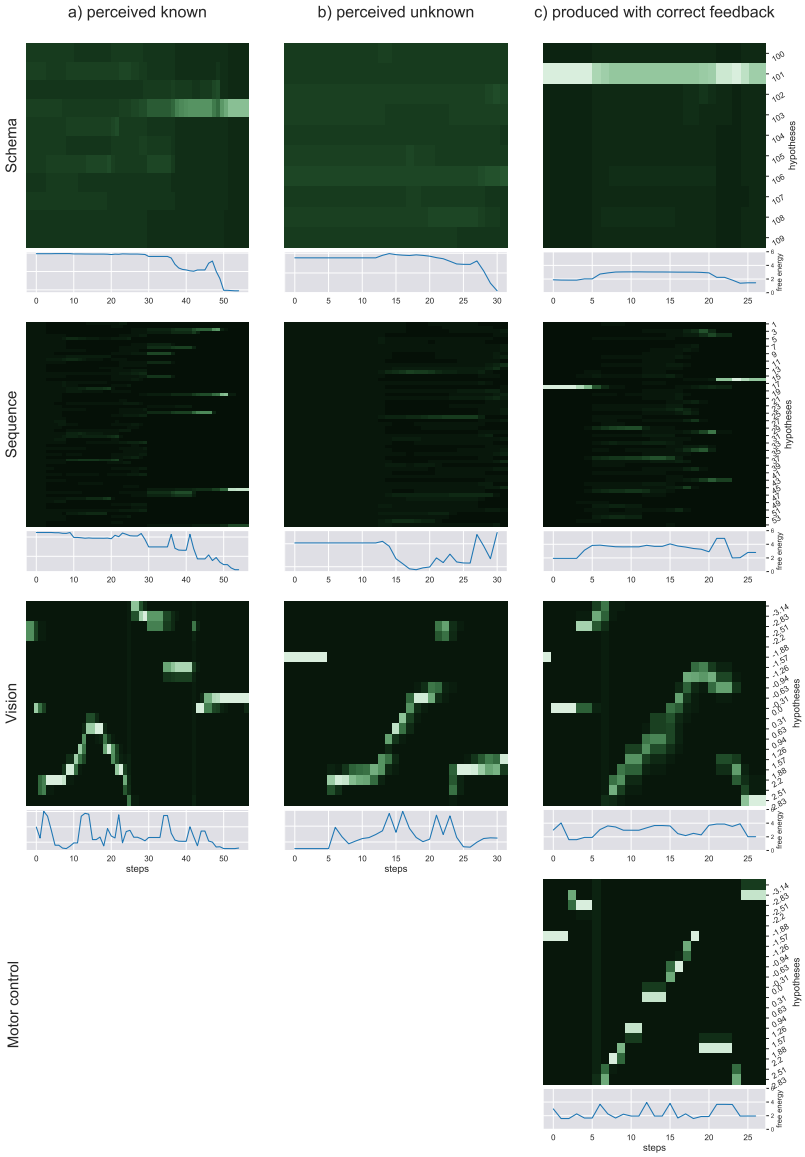


Figure 6.10: Here you can see the probability dynamics in the model as heatmaps over the probability distributions over time at the different levels of the HPBU (color coded from dark green to white, from belief probabilities $P = 0$ to $P = 0.6$ for best differentiability). The different scenarios are **(a) perceived known**: the perception of a known digit (here a 5), **(b) perceived unknown**: the perception of an unknown digit (here a 4), **(c) produced with correct feedback**: the production of a digit by means of active inference (here the digit 9). In addition the free energy dynamics for each level is drawn.

green plotted belief dynamics, a line plot of the level's free energy is drawn, showing the level of adaptation and model evidence, given the current state of the system.

To be more specific, the first column shows evaluation scenario **(a)** ***perceived known***, where HPBU faces a known handwriting example.

For this evaluation, the drawing of the digit 5 was chosen. The second column shows the model's behavior in response to scenario **(b)** ***perceived unknown***, an unknown handwriting. Here, the model will receive the drawing of the digit 4. To allow that, during training for this specific evaluation, all drawings of the digit 4 were ommitted. In the third column, the model's dynamic action production behavior can be observed as the digit 9 is produced in scenario **(c)** ***produced with correct feedback***.

RESULTS DESCRIPTION As one can see, the visual input clearly influences the perception of sequences and schemas of sequences at higher levels, thereby minimizing free energy over time. Also, during production the belief created in the schema representing the digit 9 percolates down the hierarchy, activating and acting out a selected sequence.

To be more precise, in scenario **(a)** the heatmaps show nicely how level V perceives the different movement angles over time. Simultaneously, evidence for the level S hypotheses accumulates slowly with each new salient visual feature. At first, this leads to a limited number of probable sequence representations. Finally, a single hypothesis becomes the most probable. The level C hypotheses accumulate evidence more slowly, predicting the underlying sequences. Schema level C predictions have a strong influence at the sequence level, most evident in the final distributions. There, only a number of sequences are still probable, and most of them belong the the most probable schema hypothesis.

In scenario **(b)**, evidence accumulation does not reach a necessary level to account for successful perceptual inference and level S free energy remains high. The drop at level C free energy at the end of the sequence is probably because no further evidence is received and level C can linger on its own predictions.

In scenario **(c)**, you can see that the motor control and vision level heatmaps look similar, where action production and the perception of its outcome align. At schema and sequence levels, heatmaps show how predictions are mostly met, but interestingly, evidence for the firstly predicted sequence is not met at some point. Then, another viable sequence hypothesis from the same schema hypothesis cluster becomes active after some time. This pertubation may have been caused by the spring dynamics at the motor control level.

Having a look at the free energy comparison of the different scenarios in the hierarchy's sequence level S (see fig. 6.11). This level of the hierarchy can give us an idea how well the model is able to find explanations for the perceived input, as here the schema level's higher-level expectations meet the dynamics of handwritten digits. Strong fluctuations can be a clue to highly irregular input or unreliable representations, e. g., at level V, where bottom-up sensory evidence and

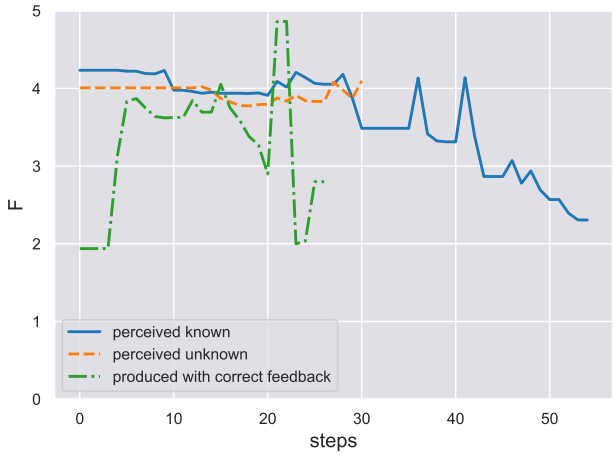


Figure 6.11: Shown are the free energy (F) plots of the sequence level S during each scenario. First, the perception of a known and unknown writing sequence clearly shows a difference in that free energy is minimized during perception of the former sequence, but not the latter. Second, the production of a writing sequence also minimizes free energy as it would in active inference as long as the sequential production is successfull, i. e., the temporal and spatial prediction of sequential acts are met.

top-down predictions can change rapidly. The perception of a known and unknown writing sequence clearly shows a difference in that free energy is minimized during perception of the former sequence, but not the latter. The production of a writing sequence also minimizes free energy as it would in active inference as long as the sequential production is successfull, i. e., the temporal and spatial predictions of action sequences are met.

To summarize, free energy is minimized as previsously expected in scenarios **(a) and (c)**, while its resistance to be minimized signifies the perception of an unknown action sequence in scenario **(b)**. The model's explanatory power in our case seems sufficient for the perception and production of sequences of writing digits. The sequence level's free energy dynamics can quickly respond to unpredicted input, but still receives predictions from the schema level C to inhibit most unlikely explanations.

6.3 DIFFERENTIATING SELF FROM OTHER

With HPBU's general ability now evaluated that it minimizes free energy during action and perception processes, we will now turn our attention towards a social perspective while – for now – staying put on the sensorimotor part. Here, the correct differentiation of self from other will be evaluated, and by that the model of sense of agency, underlying the self-attribution judgement.

The generative model that is HPBU creates action representations that spread several levels of the hierarchy, from schemas of action representations, down to simple movement primitives. In between, an action sequence hypothesis represents the occurrence of differential movements in time and space. It is these representations that allow us to take the step from abstract action to the actual details of how movement can be predicted.

When an action is produced its movements are predicted in space, and also in time. Timing information is vital to our behavior and perception, as we discussed earlier. Humans come to know their body and its effect on the environment. That is, they learn also to predict *when* the effect of their bodily action should be expected. This is the kind of information that is also predicted and evaluated in the presented computational model of sense of agency: an action's temporal and spatial attributes.

An agent's own action, predicted temporally and spatially with high precision, combined with a postdictive judgement about the agent's intention, allows to differentiate one's own from other's actions. This is the general idea underlying the differentiation of self from other, implemented in HPBU.

Here, this model is evaluted in a setting of concurrent perception and action. It is actually the *modus operandi* for the kind of active-inference based motor coordination, implemented here: during action production, the action's environmental effects are concurrently perceived both, visually and proprioceptively. To test the model's ability to infer a sense of agency from the combination of action and its effect, the visual feedback will be altered.

EVALUATION Two scenarios have been compared: In one, the production of an action was complemented with correct visual and proprioceptive feedback. This will in effect be the same action production as in scenario **(c)** of the free energy minimization evaluation (with the production of the digit 9). In the other scenario, the action producing sensorimotor part of HPBU received only correct proprioceptive

† The evaluation of sense of agency for self-other differentiation was previously discussed in Kahl and Kopp (2018).

feedback, while its visual feedback was altered (producing a 1, while receiving visual feedback for a 3).

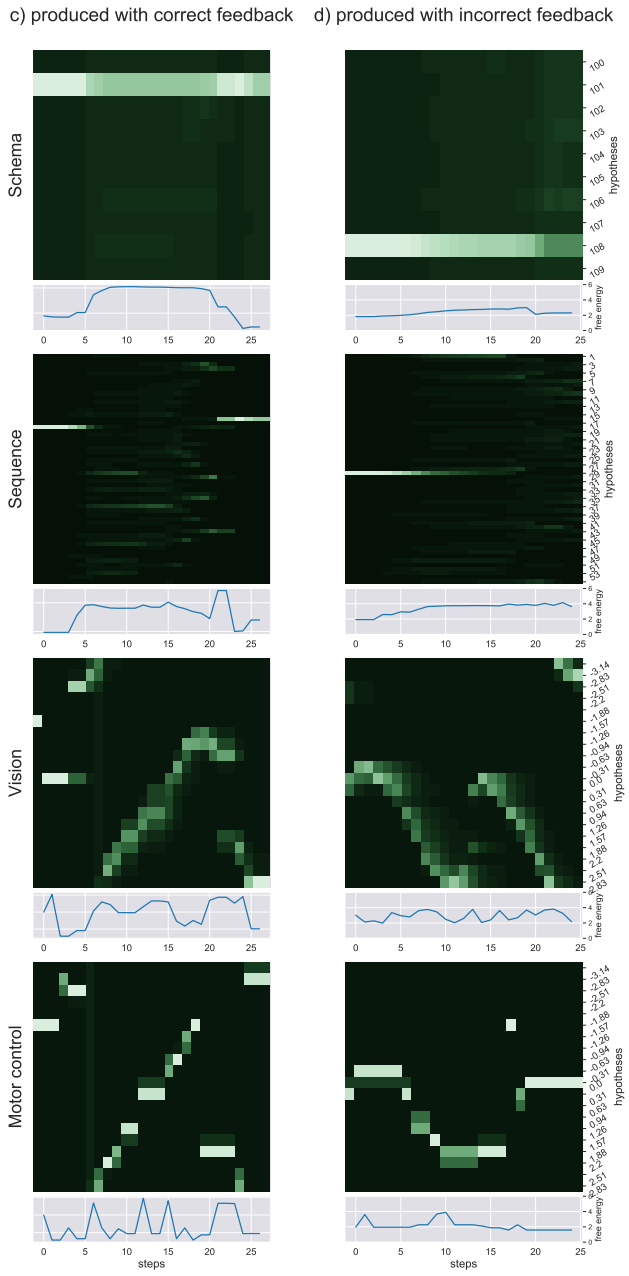


Figure 6.12: This figure is similar to fig. 6.10, showing the dynamics of two scenarios. **(c) produced with correct feedback**: the production of a digit (here a 9), with correct proprioceptive and visual feedback. **(d) produced with incorrect feedback**: correct proprioceptive feedback (of a 1) during action, while the production of another digit (here a 3) is received as visual feedback.

Similarly visualized to the free energy minimization evaluation above, here in fig. 6.12 you see two columns for each action production scenario.

In scenario **(c)** the production with correct feedback, the intention created in the schema level (representing the digit 9) percolates down the hierarchy, activating and acting out a selected sequence. In the incorrect feedback scenario **(d)**, the sequential activation is shown for producing one digit in the motor control level, while seeing the activation dynamics for visually perceiving another digit in the vision level. The resulting confusion is immediately visible at the sequence level, while at schema level, the posterior probability settles on a lower probability for the (still preferred) intended hypothesis.

RESULTS DESCRIPTION Now let us have a look at the results of the mechanism for integrating the sensorimotor sense of agency estimate over time (see sec. 5.3.1), given our simulation scenarios. The resulting SoA integration plots in fig. 6.13 show that the cue integration mechanism, supports results reported in the literature.

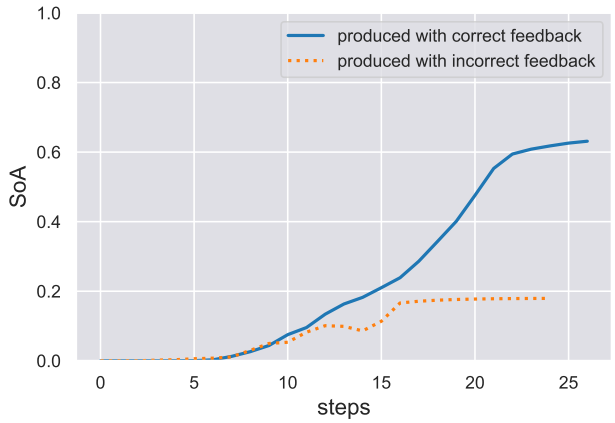


Figure 6.13: The sense of agency (*SoA*) estimate dynamics produced by the HPBU in the production scenarios of action production with correct, and the other, with incorrect feedback, respectively. In both scenarios the SoA estimate rises up to a certain point but it remains at a low level of 0.2 in the incorrect feedback scenario, where predictions of produced actions are met with contradicting visual feedback.

First, it is sound with regard to results where the precision of the predictive process was reduced and the system put more weight on postdictive processes, conforming evidence for a weighted integration based on the cues' precision (Moore and Fletcher, 2012; Synofzik et al., 2013; Wolpe et al., 2014). This aspect can be observed in scenario **(d)** with incorrect feedback, where the SoA estimate increases slowly even though a completely different digit is being perceived visually. This may be due to the fact that when drawn simultaneously, a perceived 3 and a drawn 1 start with similar trajectories, despite the roundness in the trajectory of a three. When either timing or spatial predictions are met to a degree, they can accumulate. In our simulation it is then only the lower, second curved trajectory of the 3, which is in total contrast

to the trajectory of the 1, which finally prohibits further accumulation of agency.

Second, the results are in line with a fluency effect for correct predictions of actions (Chambon et al., 2014). The accumulation of the SoA estimate over time is done using a Kalman Filter, which depends on the current free energy and precision of the sequence level S. The more accurate the hierarchy's predictions, the faster the uptake of SoA evidence (positive and negative).

Finally, even though the cue integration model is flexible with regard to the precision of predictive and postdictive cues, the scenario with incorrect feedback shows that a false attribution of SoA is not likely when both cues show no signs of agency.

Other than Friston (2011), who rely heavily on proprioceptive information, HPBU allows for visual information to solely drive motor coordination. Here, the motor coordination loop is closed using a direct connection that informs the vision level when motor control is done coordinating actions to reach a subgoal. Vision level will then check if visual information can confirm the movement and close the motor coordination loop by sending the information to level S.

6.4 MULTI-AGENT BELIEF COORDINATION

We have now seen that the sensorimotor part of HPBU has the ability to reliably recognize known handwriting sequences, and to some degree even generalize to previously unknown kinds of handwriting sequences. Its ability to act as a free energy minimizing generative model, given the uncertainties it faces, was shown during action production and perception. Also, and very much important for the next evaluation step, the model of sensorimotor sense of agency was shown to be able to correctly infer sense of agency given correct feedback. Also, it can detect confusing and unexpected feedback, and in effect prohibit the cue integration of sense of agency under these circumstances. In this work, I argue that the cue integration of sense of agency is a marker for the attribution of an action (and its outcome) to the self, rather than another agent. For that, any perceived action (-outcome) will be processed by the model of sensorimotor sense of agency and attributed to the self or another agent. For these evaluation simulations the threshold for attributing an action outcome to the self is set to 0.4, slightly biasing its attribution toward self-attributions. These attributions are represented in the person model PM of HPBU, as either me-belief (selfs) or you-belief (other).

In an ongoing interaction, these attributions will be used to compare the current set of mental states. This allows to see if work needs to be done, to get from the current state of the coordination sequence to its goal state. Another important aspect during social interaction is, whether the information provided by an interaction partner is success-

fully perceived and integrated into the agent's mental state. In HPBU, this strongly depends on the uncertainty present during perception. If uncertainty is high during perception, bottom-up information will more strongly influence beliefs. If uncertainty is low, bottom-up information might be ignored. For example, the leader agent might have low uncertainty about the follower's understanding early on, which results in a dismissal of a part of its action that might show its false belief. To counter-act the dismissal of vital information during crucial steps of the coordination sequence, we will evaluate different bias configurations of the Kalman gain K. As introduced in par. 5.2.7, it can be biased toward either preferring to maintain the top-down, or the bottom-up information.

To test the model's assumptions and find answers to the defined research questions, we need a task that allows the full model hierarchy to work within the context of free energy minimization through belief coordination. This entails that during perception processes, when known representations are evaluated, best explanations can switch and new insights need to inform the rest of the hierarchy. This basically is the work being done by the sensorimotor part. The mentalizing part itself takes this information to integrate it into the person model to track mental states of the *self* and the potential *other*. If in any coordination sequence the goal state cannot be reached successfully, another coordination sequence could be more promising to still achieve the goal state. Thus, also in the coordination sequence this switch to another model (i. e., another coordination sequence) should be able to happen.

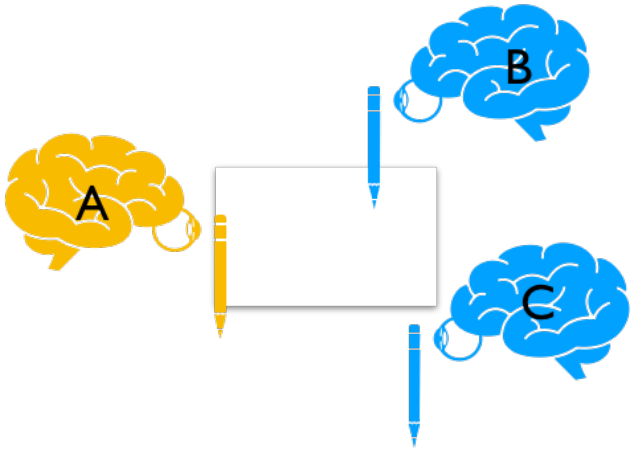


Figure 6.14: Agents are each equipped with a copy of HPBU, trained on the same corpus data. Agents differ only in their roles, with Agent A taking on the role of *leader*, while agents B and C take on the role of *follower*. The difference between them being that the leader agent will try to convince the follower agents, and make sure that they have understood, while the follower agents will primarily reciprocate their inferred beliefs about Agent A.

Another aspect to belief coordination, previously not taken into account, is the interaction with multiple agents, not only one. In such a multi-agent scenario, common ground would need to be established not only with one agent, but with all agents. For this to work, one agent would need to take care of making sure that all agents have successfully understood what was to be communicated. This is not a task for a kind of swarm intelligence, where no agent is truly in charge and all mutually strive for minimizing a measure of distance. Here, there are roles that the agents take on, with one agent taking the role of a *leader* agent during a communication game task. The other agents take the role of *followers*. You can see a sketch of the scenario in fig. 6.14. The leader agent has the task to communicate its belief of a specific digit, and make sure that the other agents have understood the correct meaning. The follower agents will perceive the interaction and will try to reciprocate, in order to convince the leader agent that they have understood the correct meaning.

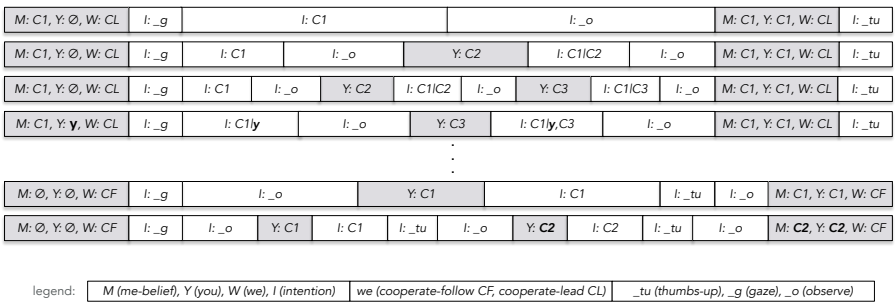


Figure 6.15: Shown are a number of coordination sequences sorted by complexity. As described in the legend, coordination sequences consist of checks for mental states (grey) and intentions (white) to either observe or produce an action as well as meta-communicative acts, i. e., thumbs-up and gaze. We-beliefs in mental state checks depend on an agent's interaction role. Intentions to produce an action refer to a placeholder for a level C representation (specified during runtime), which can involve observed false beliefs, or prior knowledge, as contrastors for sensorimotor communication.

For this to work, (as you can see in fig. 6.15), the coordination sequences for each role differ very slightly when it comes to starting an interaction and correcting for error. For example, the leader role needs a me-belief to already be set, to start its coordination sequence. Also, it is able to account for a detected false belief, trying to correct it using a sensorimotor communication strategy. The follower agents mostly just react to the leader agent's behavior, but always have the task to reciprocate what they belief they have understood. These different kinds of coordination sequences are present in all agents, but two clusters at the goal level were created. They contain the sequences that are specific to the agent's role, and the agents are configured to

prefer that role's goal state, and thus their associated coordination sequences.

One more thing needs to be added to be kept in mind: how should Kalman gain K be biased during hierarchical configurations of an intention to act or an intention to observe? Two observations have been made that show how a biased K influences the model's behavior. For perception, a high K is important to stabilize the detected prediction error, in order to give it a chance to drive the belief update at higher levels by finding better hypotheses that again, minimize free energy. During action the opposite is true, as there strong prediction errors would – in the worst case – overwrite an intended action sequence. To allow for stable action production, the top-down prediction has to be maintained, i. e., the influence of prediction error needs to be small, using a small K, biasing the belief update towards the prior.

EVALUATION We have three agents configured with the complete HPBU model trained on the same corpus. The one difference is the role with which they are configured, with one agent having the leader role, and the other two agents taking on the follower role. This configuration clamps the goal level probability distribution to that specific role-containing hypothesis which, in effect, biases the coordination sequence distributions to favor the ones suitable to the agent's role.

In this evaluation, three aspects are important: 1) The leader agent has the task to communicate its belief of a specific digit. 2) The interaction will be evaluated with regard to the leader agent's success to establish common ground with its two interaction partners. 3) The Kalman gain bias b will be evaluated with respect to its influence on an agent's ability to perceive its interaction partner's behavior, which in effect might influence the success of the belief coordination.

For this evaluation, many configurations of bias b were simulated. Here, we will compare three scenarios of the Kalman gain bias, while the intention for Agent A remains the same (producing the drawing of a 9):

- *Scenario **a*** biases K towards the prior, with a bias b of 0.3 at all levels of the hierarchy during intentions to observe.
- *Scenario **b*** biases K more strongly towards the evidence, with a bias of 0.65, throughout the hierarchy, similar to Scenario a.
- *Scenario **c*** biases K towards strongly favoring new evidence with a bias of 0.9, throughout the hierarchy, similar to Scenarios a and b.

During all scenarios the gain bias will be set to $b = 0.1$ during intentions to act.

In figures fig. 6.16, 6.17, and 6.18 you see three rows that display the belief dynamics at the schema level C, over all available schema

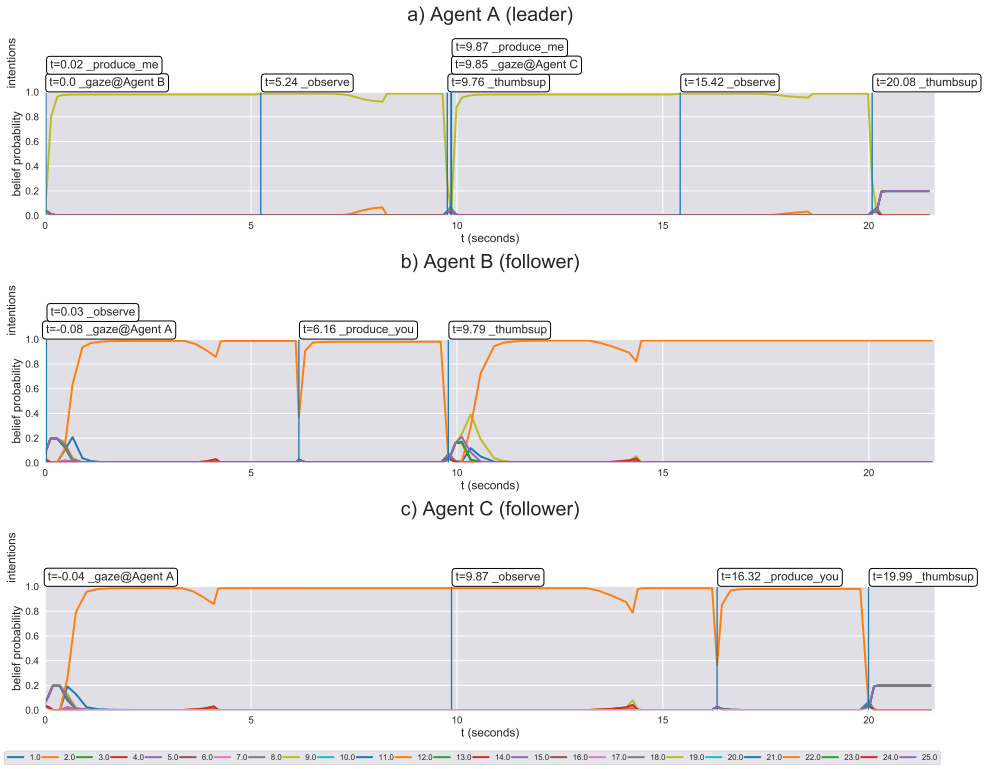


Figure 6.16: Scenario **a** with bias $b = 0.3$. Three rows display the belief dynamics at the schema level, over all available schema representations. Above each row the onset of new intentions are shown. Here, we see an example of how shared understanding cannot be established. While the leader Agent A produces its intended action successfully (drawing a 9), it does not care enough for its interaction partners. Both follower agents do not catch the correct intention behind Agent A's behavior, but this is not detected.

representations (all learned clusters of different-enough digits), which is an overall good place to look for how behavior is perceived or acted out. Above each row of belief dynamics the onset times of intentions from coordination sequence level CS are shown. Right from the start the two follower agents have the intention for *social gaze*, the meta-communicative signal. Agent A is the leader agent in the first row, who selects Agent B as the initial interaction partner, by returning the social gaze signal, establishing *mutual gaze* between them. This allows the coordination sequence to progress on.

In the following reciprocal back and forth of action and perception, over each row we see the onset timings of the intentions from the coordination sequence level. There, intentions to either *_observe* or *_produce* are triggered. The latter one also contains information about what to produce, either the me-belief or the you-belief, as it is the leader agent who will primarily produce its held me-belief, while the follower agents will primarily produce the perceived you-belief.

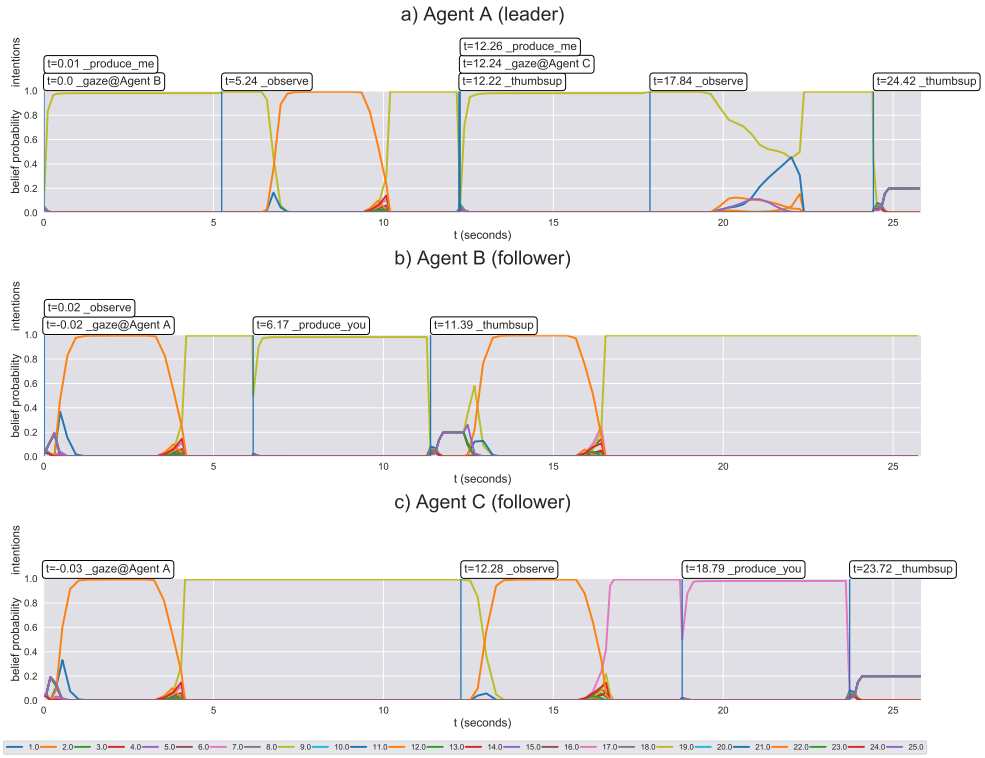


Figure 6.17: Scenario **b** with bias $b = 0.65$. Three rows display the belief dynamics at the schema level, over all available schema representations. Above each row the onset of new intentions are shown. Here, the coordination is well off with Agent B, which picks up and reciprocates the correct belief (drawing a 9). In the following coordination attempt with Agent C, its wrongly held belief is not detected by Agent A. The belief probabilities of Agent A show that Agent C's false belief is first detected but then disregarded, as Agent A's prior belief takes over again.

In the end, both agents have hopefully signaled a *_thumbsup*. Agent A will then turn to Agent C to also establish mututal gaze, because in its person model, Agent A has no common ground registered with Agent C. With mutual gaze established between them, also they can go through the belief coordination sequence, leading to common ground between all three agents. In scenario **c**, a false belief of Agent C could be detected, which triggered a switch to a more complex coordination sequence (see fig. 6.15), which allowed the attempt of repair of the false belief, using sensorimotor communication. Sensorimotor communication made use of the false belief, to select a sequence from the schema of the communication goal intention, which is maximally different than the false belief's corresponding schema. The detection of this false belief was possible due to the setting of b to a high bias towards evidence.

In order to discuss the influence of the Kalman gain bias b on the ability of the leader agent to pick up false beliefs, fig. 6.19 compares

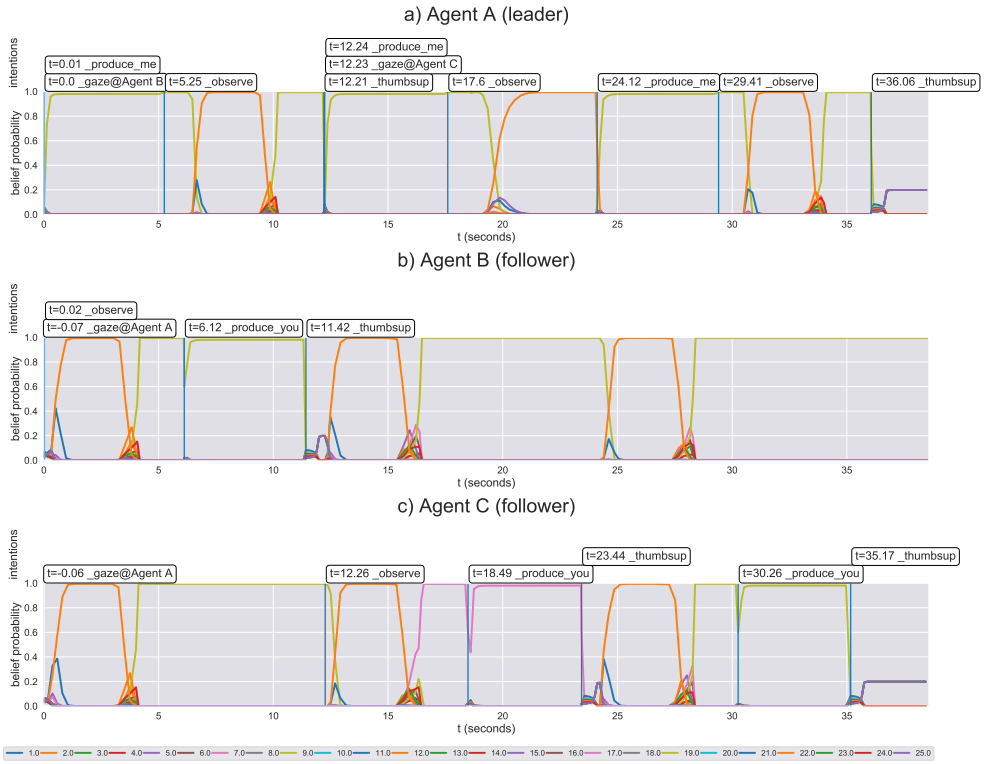


Figure 6.18: Scenario **c** with bias $b = 0.9$. Three rows display the belief dynamics at the schema level, over all available schema representations. Above each row the onset of new intentions are shown. Here, the coordination is well off with Agent B, which picks up and reciprocates the correct belief (drawing a 9). During the communication with Agent C, its false belief is detected by Agent A. Using sensorimotor communication, Agent A attempts to repair the false belief. This is successful.

the progression of belief coordination attempts depending on different biases. Displayed is the free energy F of the schema level C of the leader agent, throughout the whole interaction. Spikes in free energy mark changes at schema level beliefs and the vertical line marks the end of the interaction. The vertical line is dashed if beliefs between leader and follower agents are the same at the end of the interaction, and dotted if they are different. The interaction sequences with higher gain biases (>0.65) are longer because the leader agent is able to pick up false beliefs and attempts repairs through sensorimotor communication.

In addition, the successfull but complex belief coordination scenario **c** was chosen to see whether it also minimized free energy in the mentalizing part of HPBU (see fig. 6.20). During successful belief coordination, the actual predicted progression through the coordination sequence, one intention at a time and up to reaching the predicted goal state, minmizes free energy at the coordination sequence level. This figure (fig. 6.20) shows the complete hierarchy dynamics in the form of heat maps, during the successful belief coordination in fig. 6.18,

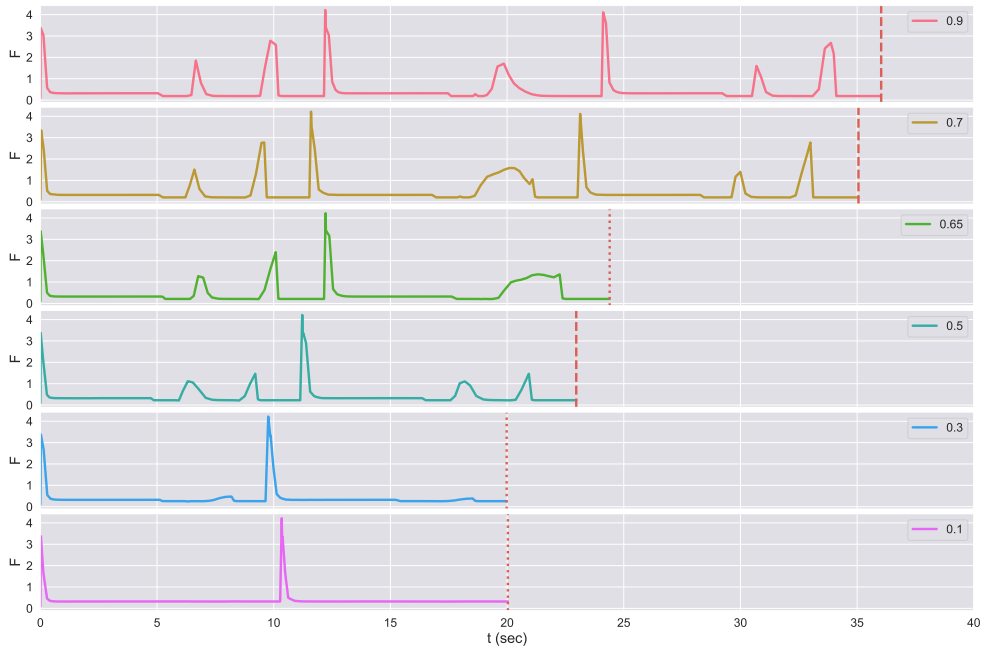


Figure 6.19: Shown is the free energy (F) at the schema level of the leader agent, plotted for different bias settings, over whole belief coordination interactions. Spikes in free energy mark changes at schema level beliefs, inferred from the follower agents. The vertical line marks the end of the interaction. It is dashed if beliefs between leader and follower agents are the same (successful belief coordination), and dotted if not.

including the free energy over time below each level's heat map. Focusing on the *CoordSeqs* and *Goals* levels, we see that free energy is indeed minimized during this successful coordination.

RESULTS DESCRIPTION Three scenarios were tested with different configurations of biasing the Kalman gain K during belief updates that is configured during different intentions from the coordination sequence level.

Generally, what can be taken from these figures of the overall back and forth of belief coordination at work, is that the mentalizing part, with its goal level and especially the coordination sequence level, constrains the sensorimotor part of HPBU. These constraints are strategically placed biases – through intentions to act or observe – permitting a flexible boundary in which it can perform its task.

The belief coordination between two agents can extend to a third agent which, when present, can be automatically included in the interaction. The role of the leader agent in this setting is vital, as it selects whom to choose as an interaction partner, based on the knowledge available in its person model.

Overall, belief coordination was possible between all three agents, even in cases of false beliefs in one of the follower agents. The detection

of such a false belief depends on the Kalman gain bias b during observation of behavior. A detected false belief triggered the switch to a coordination sequence that allowed to attempt a repair of the false belief, using sensorimotor communication. The influence of gain bias b on the agent's ability to pick up false beliefs was evaluated in different settings, showing a sweet spot for $b > 0.65$. Also, as assumed for modeling the mentalizing part of HPBU, it was evaluated that indeed, also by following through with a coordination sequence, reaching the goal state, in effect minimizes free energy.

6.5 SUMMARY

In this chapter, different aspects of the model have been evaluated. First of all, the recognition performance of the self-supervised approach to hierarchical learning was evaluated, showing a reasonable performance. Also, we saw the actual ability of the generative model to minimize free energy during action and perception. An interesting aspect here was the inability to minimize free energy, when confronted with an unknown category of handwriting.

Last but not least, the belief coordination performance was evaluated in a multi-agent scenario. We saw on the example of Agent A that the Kalman gain bias is a vital parameter, controlling the attention of the agent on the newly observed behavior. This strongly impacted the leader agent's ability to detect misunderstandings in its interaction partners. Also, we saw in the dynamics of the full hierarchy how during the successful coordination scenario free energy could be minimized at the coordination sequence level.

Now, it is time to discuss the initial research questions in light of the presented theoretical background, the modeling and its assumptions, and the evaluation results.

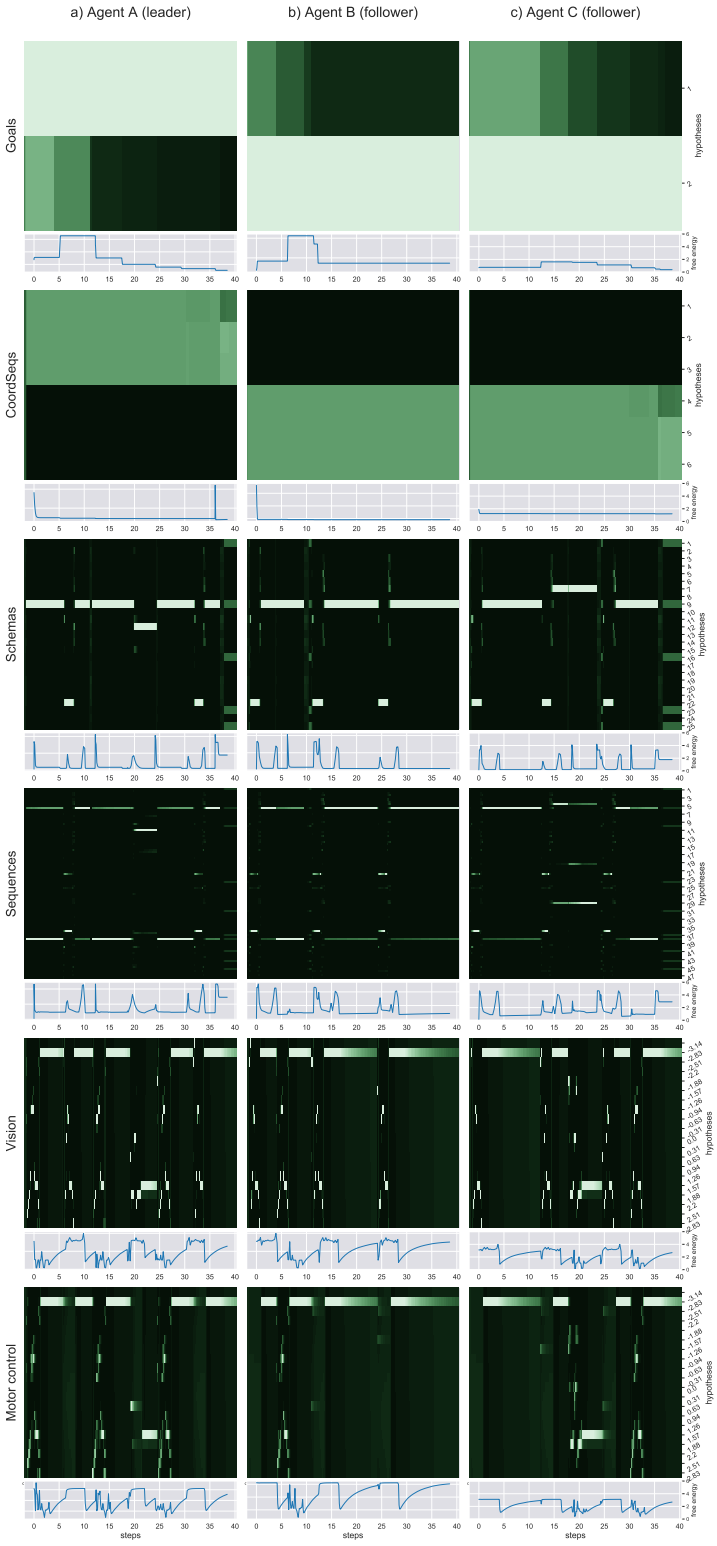


Figure 6.20: This figure shows the complete hierarchy dynamics in the form of heat maps, during the successful belief coordination in fig. 6.18, including the free energy over time below each level's heat map. Focusing on the *CoordSeqs* and *Goals* levels, we see that free energy is indeed minimized during this successful coordination attempt.

7

DISCUSSION

We as human beings master the coordination of beliefs in social interaction so effortlessly, it sometimes seems as if we had direct access to our interaction partner's minds. In the introduction I asked how the processes underlying the reduction of uncertainty in exchange with the environment, and the coordination of beliefs during social interaction, are related.

In this thesis, I have focused on the presented modeling approach, which is: to create a computational model on assumptions of predictive processing and active inference, to infer the intentions and abstract beliefs of an interaction partner, and to share them in the form of reciprocal overt behavior, to the end of achieving belief coordination.

On the way towards a computational model, the following questions needed research:

- *How are action and perception informative in social situations?* On the basic assumptions of predictive processing and active inference, a computational model of sensorimotor processing was created which, as a first step, allowed to infer and produce motor behavior. Following as a second step, based on assumptions from the literature of sense of agency and schizophrenia, a functional account of self-other distinction was created. It allows to differentiate between actions produced by oneself from those of others. This was also created on the same assumptions of predictive processing and active inference, and finally allowed to address the initial question.

- *Can active inference connect mentalizing and sensorimotor processing?* Following up on the first question, additional computational modeling was necessary to account for the processes required for mentalizing. It is also based on the same core assumptions as the initial model of sensorimotor processing. Modeling the additional mentalizing processes was necessary to perform the actual belief coordination, based on perceived behavior and performed reciprocity. Now that both processes were put on the same foundation of predictive processing and active inference, the second research question could be addressed.

Summarizing the presented computational modeling for social interaction, the following observations and inferences can be made, with regard to the research questions:

Inferrable intention representations stem from the model's ability to attennuate prediction error, finding explanations that are empirically

grounded in the information-theoretic irregularities of the input signal. These do not only cover explanations for low-level behavior, but also the interaction as a whole, up to the interaction goal.

Under the view of predictive processing, the actual perception process of inferring a belief about another's behavior becomes only necessary after failed predictions invalidate prior beliefs about that behavior. The seemless dynamic shifts of focus on explaining errors at the different levels of the HPBU model is only possible, because it does not only cover the mere necessities of symbolically representing an interaction partner's mental states. Rather, it grounds possible beliefs in the actual dynamics which become represented in the interaction with other agents, and the environment. This way, deviations from the predicted dynamics can be located at every level of abstraction, be it deviations from an action sequence, or a predicted coordination sequence.

In a way, the successful prediction of behavior and interaction goals allow for the interaction partners to bypass their perceptual loop that spans the interaction partner's perceivable behavior. Thereby, when the successful interaction is perceived from a subjective point of view, an agent effectively generates only successful predictions of itself, but for the interaction partner, and in effect the social interaction as a whole. This leaves the amount of work minimal during successful social interaction, speaking for a very efficient way of bootstrapping theory of mind. The actual *work* done in predictive processing is the model selection, necessary to attennuate prediction error. Since in successful interaction no *unpredicted* switching of models is necessary, the amount of work remains minimal. As we will discuss later, successful high-precision predictions of the social interaction, down to the behavioral levels, come close to the *feeling of direct access*, as described in the second-person neuroscience and direct social perception literature, because nothing is more direct than correctly predicting yourself.

7.1 MODELING APPROACH DISCUSSION

How have the research questions been approached in this work, which lead to these observations?

LEVELS OF ANALYSIS The information processing carried out by a human during social interaction was the focus of this presented work. Following David Marr's three level approach to understanding information processing systems (Marr, 1982), I restricted myself to finding possible descriptions at a *computational level* and a *representational (and algorithmic) level*. That is, describing the overall goal of the computation and its logic. This was first described in ch. 2, but its second part, which focused on the cognitive neuroscience underlying social interaction, overlapped with the second level of description. That is,

we first analyzed social interaction at the computational level, while already discussing cues for possible representations and algorithms of the information processing during social interaction, as exhibited by the social brain. A much stronger focus was put at this representation and algorithm level of analysis in ch. 3-ch. 4, where the background and the approach to computational modeling was described.

As said, the described computational modeling is primarily concerned with the algorithmic, representational details of the necessary computation. Of course, this omits much detail of the software implementation, where hidden assumptions may lie. Although this may be problematic on a first glance, the implementation can be described as only one instance of the computational model. It would potentially be very interesting to reimplement it and compare its simulation results. As has been discussed by Cooper and Guest (2014), one should be careful with conflating assumptions of modeling and implementation. A re-implementation of a computational model, so it could be argued, is similar to replication in helping to shine light on hidden assumptions, errors, and overlooked aspects.

With the levels of analysis in mind, what can be said about the influence of perception and action on social interaction, and to what extend can the intra-personal dynamics within the social brain be put on the basis of predictive processing and active inference?

MODELING BELIEF COORDINATION A core aspect of belief coordination is that during an interaction, it is necessary for its participants to carefully select their contributions. Following the Gricean maxims it is vital to *"[m]ake your conversational contribution such as is required, at the state at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"* (Grice, 1975, pp. 45).

In addition, understanding in human communication can be very fast and flexible, but also is prone to error. This makes it necessary for repairs to be possible. In that, reciprocity is not only a tool of belief coordination, but also helps to collect more information, either from context, extended exposure, or active repair attempts. Thus, it stabilizes a linguistic interpretation over time that at first is only *good enough*. Misunderstandings and communication errors are not fatal for social interactions, but can be repaired. As has been suggested in the running repairs hypothesis, repairs can be an integral part of communication (Healey et al., 2018). If you will, this could be described as a form of exaggerated counter-factual reasoning, i. e., taking other beliefs into account with the prior false belief as a distractor that should be avoided.

In human social interaction, it is vital that reliable understanding of other's social behavior is met by the correct prior assumptions about the interaction partner, and about the social situation.

Imagine that you find yourself on a street on your way home from work. You are approached by a person who raises her hand. You can now try to get out of the way, because you belief that the person wants to slap you in the face. Another possibility is that you recognize the person as your good friend and neighbor, who often greets you with requesting a high five greeting. The appropriate response here would be to give the requested high five.

This example shows the importance of prior assumptions (about other people and situations), and the correct recognition of behavior. In longer conversations, the process of reciprocal belief coordination becomes important, as it not only requires well-fitting prior assumptions and recognition of social behavior. It also requires the possibility to repair lingering misunderstandings, and updating those prior beliefs about interaction partners. The process of belief coordination presented here is similar to the *we-mode*, proposed by Frith (2012): An implicit form of mentalizing that bootstraps the attribution of mental states, which here allows to reciprocally update common ground, while contextual information and prior information can influence behavioral understanding top-down.

It is the presented combination of interpreting non-verbal behavior during social interaction that underlies the title of this thesis: "Social Motorics". That is, a combination of an embodied grounding of processes of perception and production of action, in dynamic interaction with higher-level abstractions of subsequent events within a social context.

DEEP TEMPORAL MODEL The computational modeling approach presented here solves the necessary abstractions and dynamics, of uncertainty reduction during social interaction, using a hierarchical approach of updating prior beliefs over longer temporal time scales. The computational cognitive model is called HPBU, and is based on a hierarchy of generative processes. These not only allow the prediction of spatial relations in the perceived behavior, but also temporal relations.

One basic assumption is that the dynamic environment (which includes the agent's behavior) unfolds as an ordered sequence of states, where input can induce dynamic trajectories in representational state space, embedded in the model. HPBU is a generative model of such trajectories, but in a hierarchical form that enables it to categorize and represent similar sequences. Such a hierarchical generative model can exhibit multiple time scales, i. e., sequences covering longer time scale can generate sequences of shorter time scale. In other words, at different levels of the hierarchy generative processes allow to predict sequences of events over time. Together, this hierarchy is an example of a generative model that infers nested sequences of state transitions

of shorter temporal stretches within sequences of longer temporal stretches. This has previously been described as a so-called *deep temporal model* (Friston et al., 2017b).

HPBU as it was described, presents these properties: it infers sequences of belief coordination, in which parts of a coordination sequence contain nested sequences, either of action understanding or action production. In its properties of an uncertainty minimizing probabilistic model, it can be described as a local search at every level of the hierarchy for inferring the approximate posterior, given the message passing that over time reaches the level in the form of bottom-up input and top-down predictions. Each level at every time step, when a new input or prediction is available, employs a form of variational belief updating. That is, it minimizes free energy by finding available representations (or models) that better predict the evidence (or maximize model evidence), before predicting the next time step from the approximate maximum posterior representation. The sequence-processing levels calculate their likelihood based on recent salient movements, collected since the completion of the last sequence.

Further in-depth treatment of the presented specialized algorithm's formal comparability to similar approaches will not be part of this work, but should be part of a possible future extension toward a generalized machine-learning framework.

7.2 EVALUATION DISCUSSION

With the general theoretical background and the modeling discussed, what did the evaluation reveal?

SENSORIMOTOR EVALUATION DISCUSSION The recognition performance of HPBU's self-supervised learning approach has been evaluated, with generally good performance, already after training on one user source, with slight increases in classification and generalization performance after repeated training on that source. The best performance could be achieved through a pooling of user sources to train on, but this led to a very high number of necessary representations, which are computationally demanding.

Increases in recognition performance could possibly be achievable through optimizations that would be needed to go beyond representing mere information-theoretic irregularities. These include the representation's reliability in relation to other – already known – representations. What comes to mind is, to account for the frequency of a representation's success in reducing free energy. In addition, the possibility of a life-long learning approach could be considered, with a representation not only of reliability, but also with an account of frequency of encounters, over the model's runtime.

Generally, the presented corpus of handwritten digits seem to contain exemplars that are hard to distinguish. Especially, since even with the highest number of representations two digits where hard to distinguish for the model. The sequence comparison measures seemed to be insufficient to account for all exemplars of the digits 0 and 6.

We have also seen results from a master's thesis that compared HPBU to state-of-the-art neural network approaches, which could reach the initial level of classification performance after 50 epochs.

The model makes judgements about the suitability of a novel sequence to become represented, and to be assumed to be dependend upon one of several hidden variables (e. g., the schema clusters). That is, it may become a member of a cluster, or trigger the creation of a new cluster representation. This ability to extend its state space is what sets HPBU apart from traditional modeling approaches. To my knowledge, no standard approach to hierarchical Bayesian modeling allows for these kinds of dynamic updates that allow to add new representations on the fly, in effect extending the discrete probability distributions.

Another focus of the evaluation was to see if HPBU actually minimizes free energy during action and perception, after it was trained. Here, we could see that although being able to minimize free energy during the perception of a known handwriting sequence, it did not minimize free energy when confronted with an unknown handwriting sequence. This shows the ability to differentiate between a known and an unknown signal. Also, it exposes the point at which learning of new sequences takes place, with the possibility of either an extension of existing schema clusters, or a rearrangement of sequences into additional schema clusters.

The inferred sensorimotor sense of agency during action is central to the correct attribution of beliefs of perceived action. It strongly depends on the parametrization of the gain bias b (to be discussed soon) and the context of the coordination sequence – whether the agent intends to act or observe – and highlights the uncertainty immanent in the sensorimotor loop. We see in the evaluation of an action with correct visual and proprioceptive feedback how SoA can successfully be established in temporal and spatial predictions, and both feedback types. When one of these predictions is not met, as in the second evaluation, sensorimotor SoA cannot be inferred to the level of the first evaluation. That is, correct proprioceptive feedback was combined with incorrect visual feedback. This way, a situation has to be simulated where, during the own action production, another agent's behavior is visually observed. This quite nicely demonstrates also the dual use of the sensorimotor system of perception and production processes. During the correct – and to some degree automatic – production of an own action sequence, the perception of another's

action could potentially still be inferred, but it is ignored at the moment. What was missing is a mechanism to actually *cancel* a failed action production when the evidence suggests unpredicted action effects. Still, a prediction of this model is that during these dual-use cases, the differentiation of self from another agent's action cannot be guaranteed, if the perceived and produced behavior is too similar.

GENERAL MENTALIZING EVALUATION DISCUSSION Belief coordination was evaluated in a simulation of a multi-agent communication game, with one agent having the leader role and the other agents that of followers. The leading agent would try to make sure that the followers have correctly understood its communicated intention. The sequence of coordination attempts could be shown to be able to lead to the establishment of common ground between all agents. Three scenarios were tested with different configurations of biasing the Kalman gain K during belief updates that is configured during different intentions from the coordination sequence level. What can be said about the belief coordination in these different conditions of biasing the uptake of different beliefs?

In fig. 6.16, which shows the coordination of *Scenario a* (using a bias $b = 0.3$), we see an example of how shared understanding cannot be established. While the leader agent produces its intended action of schema *9* successfully, it does not care enough for its interaction partners. Both follower agents do not catch the correct intention behind Agent A's behavior. Also, Agent A does not detected their false beliefs.

The coordination goes similarly in *Scenario b* (using a bias of $b = 0.65$), where Agent A again attempts at communicating schema *9* (which stands for writing a 9). Here, the coordination starts off well, where the follower Agent B picks up and reciprocates the correct belief (drawing a 9). In the following coordination attempt with Agent C, its wrongly held belief is not detected by Agent A. The belief probabilities of Agent A show that Agent C's false belief is first detected but then disregarded, as Agent A's prior belief takes over again (please see fig. 6.17).

Finally in *Scenario c* (see fig. 6.18), we get to see a successful communication attempt, despite false beliefs. Here, the leader Agent A puts enough attention on the interaction partner's behavior, using a bias of $b = 0.9$. Again, Agent A has the intention to communicate schema *9* and the coordination is successful with Agent B. During the communication with Agent C, its false belief is detected by Agent A. The false belief triggers the switch to a coordination sequence that allows to attempt a repair of the false belief, using sensorimotor communication. This is successful.

First of all, it could be shown that belief coordination can be successful in establishing common ground. When belief coordination was not successful, mostly the inattentive leader agent was to blame for

not detecting misunderstandings. Attentiveness was parameterized by applying a gain bias b to the belief updating performed at every level of the respective agents' hierarchies. The analysis of the influence of different parameter settings for b showed a sweet spot for picking up false beliefs and repairing them (see fig. 6.19). Despite false beliefs in follower agents, only biases of 0.7 and higher allow for successful repair attempts. These biases allow for false beliefs to be perceived and processed to trigger attempts for repair. A bias greater than 0.65 seems to be the sweet spot for the detectability of false beliefs. With a bias of 0.65, the increase in free energy at t=20 might signify the detection of a false belief, but it is not strong enough for the coordination sequences at the next higher level to trigger a repair attempt (see spike at t=23 in bias 0.7). In the bias setting of 0.5, beliefs between agents are the same, which might be due to chance. Generally, the analysis shows that the gain on prediction error influences free energy due to the Kalman gain bias increase.

The Kalman bias parameterization that led to unsuccessful belief coordination can be compared to the so-called confirmation bias, as observed in humans. The confirmation bias has first been described by Wason (1960), as a bias that during an inquiry we seem to ask questions that seem to confirm our hypothesis, rather than trying to disprove it. Similarly, this has been described in the light of the cognitive bias called availability heuristic (Tversky and Kahneman, 1973), where they observed that people tend to bias their predictions on things that are salient (likelihood of causal influence) or vivid (easily recalled and convincing), rather than accounting for its probability.

INTENTIONS AND INTENDED ACTION A possible line of criticism is the long-range connection for the intention signal, which is perculating the hierarchy in order to configure the sensorimotor hierarchy part to the needs of the mentalizing hierarchy. E. g., the intention signal is received by level C (or schema level) and tags one of its hypotheses for production in active inference. Then, using additional long-range connections to other levels in the hierarchy the intention is spread, giving an initial boost of probability to associated hypotheses at those levels. In the present implementation the intention signal is also maintaned to continuously tag the intended action. This way, during inference, sudden switches to other representations (e. g., other probable action sequences), are inhibited.

Similarly, activity is maintained during attentional tasks in area MT (Treue and Martinez Trujillo, 1999). There, monkeys were tasked to follow a single moving visual stimulus while other movements were also on display. The maintained activity only decreased once the task was performed and the stimulus vanished.

One should naturally assume that the probability of intended action representations should automatically be maintained by successful

predictions of actions. In reality, in the model, small differences in the feedback to early actions of an action sequence can lead to increased likelihood of other, at that moment more similar sequence representations that not always belong to the intended schema. It may well be the case that in an extended hierarchy, the maintained tag on the intended representation may be provided by higher levels through appropriate priors.

For successful active inference, incoming sensory signals to some degree have to be ignored in order to be able to initiate action. As a sidenote, such a necessity to maintain the intended action during production was also identified by Doiega (2018), who discusses a predictive-processing interpretation of *M-autonomy*, i. e., a formulation of mental agency for intentional action (Metzinger, 2015).

IMPLICATIONS OF THE PRECISION WEIGHTING BIAS In fig. 6.19 we have seen how a bias on Kalman gain K could strongly influence the gain of prediction error on free energy throughout the model hierarchy. Thus, in addition to the intention signal that configures the sensorimotor part of HPBU, also the precision weighting was biased in order to make the intentions to act or to observe happen. A weak bias to the influence of prediction errors could lead to uninterrupted action and observation, without sudden switches to similar, but unintended representation hypotheses. A strong bias can lead to increasing fluctuations in free energy, and repeated switching between similar representations, while reducing the chance to fixate on possibly false representation hypotheses.

This strategically applied bias b of the application of the calculated Kalman gain K during belief update seems to be a key parameter. Specifically in the task of belief coordination, if not properly set, the evidence for falsely held beliefs of an interaction partner can be overlooked.

Precision weighting has previously also been associated with the functional role of dopamine at the synapses, to balance bottom-up sensory information with top-down prior beliefs during hierarchical inference (e. g., Friston et al., 2012). In simulations they show how dopaminergic lesions can produce behavior similar to neurological disorders, such as Parkinson's disease. If anything, this at least highlights how central the role of *integration* of prior beliefs and sensory evidence is, in systems faced with uncertainty. More work is definitely needed to properly understand how, when and which bias at the different levels in the hierarchy should be applied, and what that bias depends on.

Also, the found sweet spot could only be reproduced for a handful of representations. It is not universal for all representations. Further analysis of the underlying dynamics is here ommitted, but should be part of a thorough investigation in future work. Hypothetically,

it could allow for the strategic placement of attention on points of inquiry, either within the hierarchy, or in the external environment, as perceived and influenced through the sensorimotor system. To make the point more explicit: a strategically applied bias, *placed by the system on itself*, is a form of *meta-cognition*. Sadly, it is beyond the scope of this work to discuss the implications in general, of systems that place strategic biases on their own integration of prior beliefs with sensory evidence. What has been shown here is that its correct application is vital for the successful coordination of beliefs during social interaction.

To summarize: the correct integration of prior beliefs and sensory evidence is vital for the process of approximating correct posteriors at the levels of HPBU. Also, the presented treatment of the intra-personal dynamics within the social brain has implications for the understanding of successful direct social interaction.

IMPLICATIONS FOR DIRECT SOCIAL INTERACTION Within the presented model of HPBU, active inference and predictive processing connects mentalizing and sensorimotor processes. This is done in a way that may account for the subjective feeling of direct access to our interaction partner's minds. First of all, it allows for beliefs about an interaction partner to act as a prior for the recognition of behavior. Thereby it influences the likelihood of the perceived understanding of an action (remember the example about the high-five greeting neighbor). Also, it allows to correct for errors in understanding, and to update beliefs over time. Another integral aspect is the precision weighting that can bias belief updating toward favoring a stable prior belief, effectively ignoring prediction errors. If precision weighting is biased correctly, reciprocity becomes possible, which can lead to a shared understanding between interaction partners. Such shared understanding would be realized by a prior and a precision weighting that entails good predictions about each other's beliefs and intentions. In effect, this prior along with the correct precision weighting helps to minimize free energy at all levels of the hierarchy, by allowing to efficiently anticipate, react and sometimes ignore prediction errors that might occur in the perceptual loop, including the observation of an interaction partner's behavior. This process can be understood as a form of bypassing the conscious error-correction necessary for inferring each other's beliefs and intentions, thus again making necessary adjustments automatically and unconsciously (remember Helmholtz's unconscious inference). This bypass allows each interaction partner to assume each other's beliefs without question, which may result in the feeling of direct access to the other's mind. In other words: the subjective feeling of direct access to an interaction partner's mind is due to continued correct attennuation of errors to predictions about the other's behavior, during social interaction.

In summary, this can be described as a hypothesis for a mechanistic account, which is missing in the dark matter of social neuroscience (as described in sec. 2.2.3). Especially the observed impairment in HFA patients could potentially be shed some light on. There, the implicit process in direct social interaction is impaired which, in neurotypical people, allows them to automatically reorient themselves and integrate social cues. This lack of automaticity during implicit mentalizing is described to be overwhelming for HFA patients at times. Especially, when directly engaged in interaction, in contrast to the patient being a passive observer (Schilbach et al., 2013).

A prediction that results from this hypothesis, is that in HFA patients the attennuations, necessary for implicit and automatic processing, are to some degree not possible. From the literature on impaired sensory attennuation that is suggested to underlie positive symptoms of schizophrenia (Adams et al., 2013; Brown et al., 2013; Friston and Frith, 1995; van der Weiden et al., 2015) to the prediction hypothesis of autism (de Cruys et al., 2014; von der Lühe et al., 2016), there are pointers toward the idea of a spectrum of incorrectly tuned precision weightings. Possibly, there is an over-reliance on *bottom-up* information in schizophrenic patients and an over-reliance on over-generalized *top-down* predictions in autism patients. For HFA patients, the automatic processing would often not be achieved, because the possible attennuations are not detailed enough for the automatic processing to occur.

To highlight this again, the shown integration of the sensorimotor part with the mentalizing part suggests an intra-personal dynamic that influences the inter-personal dynamics of belief coordination.

8

CONCLUSION

Let us now conclude this thesis with a brief summary of its main arguments, results and contributions (sec. 8.2). We will also discuss their implications as well as the limitations of this work, and possible future research directions (sec. 8.3).

8.1 OVERALL SUMMARY

In this thesis I sought to address the problem of how interaction partners can come to a shared understanding using a predictive processing based mechanism for the interplay within the social brain. Entailed in this problem are the two main research questions that motivated the work for the thesis. To tackle this problem, two of Marr's levels of analysis were applied to differentiate the computational from the funtional and algorithmic levels. The computational level was used to analyze the modeling problem and identify the necessary processes underlying the task of achieving shared understanding in social interaction. In the next step, the functional accounts and computational modeling were approached at the algorithmic analysis level.

The first research question was: *"How are action and perception informative in social situations?"*. In the context of this question we have visited and discussed aspects relevant for understanding social interaction in general. Then, we focused on belief coordination, non-verbal communication and the different findings from conversation analysis. They uncover the core problem people face when they try to establish shared understanding with one or many interaction partners. Mainly the problem is one of uncertainty, where the inferred understanding of exchanged communicative signals being subject to influence from many sources. One influence comes from past experience and context, regarding the information itself. The other is the social influence from past experience with the interaction partner. Both can create good-enough lingering understandings which, through the belief coordinating process, can be tested and repaired, when necessary.

Reviewed accounts of computationally modeling all missed one or more of the elements necessary for dynamically perceiving and production behavior during communication, to the end of coordinating inferred beliefs in direct social interaction (see sec. 3.6). Thus, the computational model called HPBU was created based on predictive processing. It has uncertainty at its core, with a central strategy of decision making that is setup up to minimize this uncertainty (or free energy) on the different levels of its hierarchy. Also, it handles inferred

beliefs as possessing causal powers that can be employed and tested in the form of predictions. Here, the levels of the hierarchy represent increasing abstractions (sequences and schemas) over visuo-motor movement primitives. In a form of active inference stable schema-level beliefs can drive its lower-level sequence and motor levels to generate predicted actions in order to minimize free energy. At the same time, schema-level beliefs are influenced by perceived movement sequences, minimizing free energy by driving its higher levels to select best-fitting representations.

One focus of evaluation was to see under which circumstances the model actually minimizes free energy. Results of these evaluations show that free energy is minimized during action and perception of known representations, but not if the observed behavior is unknown or if feedback during action production is unpredicted. Thus, the model makes judgements about the suitability of a novel sequence to become represented, and to extend its state space. As reviewed, this sets HPBU apart from traditional modeling approaches. It was found that the model allows not only to handle uncertainty during processes of perception and production of communicative signals. It can also use predictions about the embodied nature of its learned representations, which culminated in a functional account of SoA for self-other differentiation. We see in the evaluation of an action with correct feedback how SoA can successfully be established in temporal and spatial predictions while, when predictions are not met, as in the second evaluation, the inferred value for SoA stays low. Thus, the applied account allows HPBU to differentiate its own actions from those of its interaction partners.

We also visited and discussed the social cognitive neuroscience perspective on the problem of achieving shared understanding. There are significant differences in how the brain processes information with or without an interaction partner. Only with an interaction partner prior information about the interaction partner is taken into account, and only then does the perceived behavior become informative with regard to a belief-coordinating process.

The second question was: *"Can active inference connect mentalizing and sensorimotor processing?"*. Still missing was a mechanistic account about the interplay of the two functional networks of the social brain, involved during social interaction. Here, such a mechanistic account of mentalizing and sensorimotor processing was put forward in the form of a computational model; an extended HPBU hierarchy that forms a combined social predictive processing hierarchy. In it, beliefs about an ongoing social interaction are stored in the form of mental state attributions that inform state-goal pairs of mental states that track the success of an ongoing belief coordination attempt. Interaction goals, inferred using these structures, are successively fulfilled using coordination sequences that employ strategies for reciprocal belief

coordination. These can influence further processing in the form of strategically applied intentions to act or perceive. Those intentions are effectively biasing hierarchical belief updates to either focus on, or ignore new evidence. By biasing belief updates and through top-down predictions the mentalizing part influences the sensorimotor part of the hierarchy during action production and perception. This is a form of active inference that allows the mentalizing part to test its beliefs in a form of reciprocal belief coordination. Also, the top-down influences were set up to allow for an action sequence selection to be informed by false beliefs attributed to a specific interaction partner. This is a form of sensorimotor communication that allows for a *recipient-optimized* communication, making the overall reciprocal communication strategy more efficient.

The applied model was evaluated using a multi-agent communication game, with one agent having the leader role and the other agents that of followers. The evaluation was focused on evaluating the influence of biasing the Kalman gain K during belief updates, as strategically configured during different intentions from the coordination sequence level. Testing the model's viability to coordinate beliefs under these conditions, evaluations found that shared understanding can successfully be established. The Kalman gain bias b needs to correctly be parameterized to a level that allows for a form of confirmation bias to be overcome that is specific to the employd communication goal; here in the form of a specific digit. This is needed in order to successfully integrate information during the different phases of belief coordination, identify false beliefs, and for repair strategies in the form of sensorimotor communication to succeed.

In conclusion, the presented work sheds light on the importance of handling uncertainty when interacting with the environment – or specifically – during reciprocal belief coordination with other people. This was possible due to significant strides in computational modeling of motor coordination, and in enabling computational embodied models of the social brain to engage in multi-agent interaction during communicative settings. The most important contribution is the mechanistic account of the interplay between mentalizing and sensorimotor processing, with implications for the notion of subjective direct access to other's minds during social interaction.

8.2 CONTRIBUTION SUMMARY

I regard the work presented in this thesis to be relevant to ongoing discourses within multiple fields of research.

COMPUTATIONAL COGNITIVE MODELING With the current focus on deep learning approaches to machine learning and computational

cognition, I think it is necessary to contrast the presented work's contribution to the field of computational cognitive modeling.

First of all, HPBU represents a different approach to learning than weight updating of existing connections. Rather, it accounts for uncertainty in the input signal by "learning" new sequences on an account of information-theoretic irregularity. In effect, the model extends its sequence representation repertoire in a self-supervised approach. This sequence gets abstracted upon in the hierarchy, thus it becomes embedded in clusters of similar sequences, upon which further abstractions of sequential processing can be performed, if necessary. To my knowledge, this is the first account of a hierarchical probabilistic model to handle uncertainty in this way.

Doing so, HPBU goes beyond conceptual models of hierarchical processing, e. g., of motor coordination (Wolpert et al., 2003). It not only handles uncertainty during coordination, but can learn new sequences if necessary. Also, it goes beyond modeling of coupled linear oscillators (e. g., Dumas et al., 2012a), in that it does not adapt the model dynamically to a single state. Rather, it handles uncertainty during sequences of events, continuously finding minimal free energy states.

The actual dynamics within HPBU more closely resemble what has been described by Friston et al. (2017b) as a deep temporal model. At different levels of the hierarchy, generative processes allow to predict sequences of events over time. Together, the hierarchy is an example of a generative model that infers nested sequences of state transitions of shorter temporal stretches within sequences of longer temporal stretches.

In caring also for the temporal parameters of its embodiment, HPBU needed to go *beyond* the original idea of a deep temporal model, as it not only allows to have nested sequences of state transitions on different clock speeds. Rather, the represented temporal dynamics of the state transitions themselves are parameterized from experience. These learned temporal transitions allow the model not only to predict *what* to expect, but also *when* it is to be expect.

A MECHANISTIC ACCOUNT OF THE INTERPLAY IN THE SOCIAL BRAIN This thesis also adds to the body of work on social neuroscience. For long, a mechanistic account for the interplay between sensorimotor processes and mentalizing processes was missing. One that could account for the differential activation in imaging data between participants in social situations. The missing mechanism was dubbed "the dark matter of social neuroscience" (Przyrembel et al., 2012).

As a step toward a complete mechanistic account of the social brain, the presented work contributes a computational model of sensorimotor sense of agency. It allows for a distinction of own actions from

that of others. This model is based on functional mechanisms identified in the literature, spanning ideas from the comparator model, to a proposed account of disturbed precision encodings at the core of the Schizophrenia pathology. Combined, this enables a self-other distinction that feeds into the mentalizing part of HPBU, informing further processing in the context of social interaction.

The mentalizing part of HPBU was exemplified in scenarios of belief coordination during social interaction. The presented model contains a hierarchical structure, based on state transitions between mental state representations, which could also be described as social affordances. Similar state transition sequences (or coordination sequences) towards a communicative goal of a belief coordination are clustered. This allows for different approaches to the same problem of belief coordination, with or without means for repair. Mental state representations receive information from the sensorimotor part of the model. With that information the mentalizing part can track the belief coordination, and configure the sensorimotor part to achieve its communicative goal. The sensorimotor part either observes an interaction partner's behavior, or produces communicative behavior.

Engrained in this mechanistic account of the social brain is that HPBU is based on predictive processing and active inference. This allows for prior information about an interaction partner to influence the likelihood in belief update processes, during perception or production of behavior. Reciprocity allows to correct for errors in understanding and updates beliefs, so that over time, a shared understanding between interaction partners can be established. Shared understanding (in the form of minimized free energy at sensorimotor levels of the hierarchy) effectively allows to bypass the perceptual loop that includes the observation of an interaction partner's behavior. This allows the agent to only predict itself (at mentalizing levels of the hierarchy), as long as the prediction errors from the interaction partner's behavior can continuously be correctly attennuated.

An interesting implication of this view is that this bypass allows each interaction partner to assume each other's beliefs without question. In humans, such a similar process may result in the subjective feeling of direct access to the other's mind, since nothing is more direct than observing your own thoughts.

Another important contribution, though not only for the field of social neuroscience, is the importance of precision weighting. Precision weighting, in the form of a biased Kalman gain, was found to be vital for the balanced updating of beliefs, which we have seen to impact not only the correct attribution of sense of agency to own actions. We have also seen it to be vital for a social agent to gain access to its interaction partner's beliefs, by observing their actions without its prior beliefs overwriting the new information during the inference process.

The impact of precision weighting on the pathology of Schizophrenia, e. g., in the form of attennuation, has already been discussed, as it directly influences also the attribution of agency and the understanding of an active self. For autism it has also been discussed that predictive processes might be at the core of the automatic processing underlying implicit mentalizing, as seen in HFA patients. To my present knowledge though, the role of precision weighting has not been highlighted to the same degree in disturbed social cognition, as it has been for cases of positive symptoms of Schizophrenia.

Clearly, the role of precision weighting should more deeply be investigated. Thus, let us now come to discuss the limitations and the possible outlook for the presented body of work.

8.3 LIMITATIONS AND FUTURE WORK

With the many assumptions that carry the presented modeling approach, there are some that need to be questioned, if the model's conceptual power is to be evaluated and extended further.

One limitation that is quite obvious, is that the HPBU agents did not play their communication game with an actual human participant. Much of the modeling work presented here would not have been possible without the empirical findings in the literature on social cognition, conversation analysis, and social neuroscience. Thus, a true interaction with human participants, as the de-facto gold standard of social agents, should be a primary milestone for future development. One problem with this is the open question of finding the correct parameterization for precision weighting during an interaction as such. Precision weighting should be another focus of future developments of this model. In the presented model belief updates were heavily influenced by an uncertainty based Kalman gain K, which again was influenced by a gain bias b. Only few gain bias parametrizations were successful for a specific digit to allow for successful application of repair strategies and the detection of a false belief. Not all performed parametrizations were included in this work, but I hypothesize that the success of this gain bias parameter is dependent on the kind of digit and variability of its representations. To make the point again: a strategically applied bias, *placed by the system on itself*, is a form of meta-cognition, which could be learned. Hypothetically, it could allow for the strategic placement of attention on points of inquiry, either within the hierarchy, or in the external environment, as perceived and influenced through the sensorimotor system.

A first approach to mediate this problem would be to use other training corpora that would allow to find more robust representations. Using more extensive corpora would probably lead to increasing necessary computational resources. To handle these, a good next step would be to make use of parallelization on GPUs, by performing

the necessary likelihood calculations in parallel. Also, a coupling of connectionist accounts of processing sequential information, like recurrent neural networks, with a predictive processing based account, should be considered to make use of the strides in performance of these accounts.

Another possibility to tackle this problem would be to only *compare* human task performance with that of the model in specialized tasks. For example, the presented mechanistic account of the interplay within the social brain could be investigated in the future, by comparing the model performance with that of HFA patients. For that, the model's precision weighting could then be biased to match the performance of social information integration of HFA patients. Similarly, for patients suffering from positive symptoms of Schizophrenia a comparison also with the model's performance on a specialized task could be performed.

If such an endevour would be successful and a parameter spectrum of precision weighting could be established, an improved social interaction, with a balanced agent, would become possible. Also then, the repair mechanism embedded in the hierarchy would become more effective in making use of *sensorimotor communication*. This would allow to reciprocate during belief coordination in a way that strategically takes the other agent's belief into account more extensively.

On a similar note, Brandi et al. (2019) propose the concept of social agency, to refer to the experience of agency in a social interaction. They develop a mechanistic account for social agency, based on predictive processing, and propose to test it on patients with different disorders, to see how it impacts social agency. The proposed mechanistic process is based on a hierarchical representation of social interactions. I belief the account of sense of agency, already available in HPBU, could provide a measure of social agency when applied at the level of social interaction, i. e., HPBU's coordination-sequence level.

Another limitation of the presented work is the depth of the treatment of the model's mathematical validity and suitability as a generalized framework. As discussed, HPBU combines properties of an uncertainty minimizing hierarchical probabilistic model, with properties of a linear dynamical system. It employes a local search at every level of the hierarchy for inferring the approximate posterior, given the message passing that over time reaches the level, in the form of bottom-up input and top-down predictions. Each local search has the form of a variational belief updating (see par. 4.1.4) which, by selecting a better-fitting model, minimizes free energy. From the updated approximate maximum posterior representation, a better prediction can be made for the next time step. As was previously proposed, deep temporal models can be expressed as a hierarchy of Markov decision processes (Friston et al., 2017b). Although, this example does not cover

explicit representation – and thus predictability – of temporal relations, or learning of new representations.

Self-supervised learning in HPBU is currently restricted to the extension of the state space by adding new representations at the sequence and schema levels. In future developments the present learning approach should be taken a step further, maybe by optimizing the connection weights between clusters and their associated representations. Such a future development would allow to learn the comparison functions used to calculate the likelihoods, which at the moment are static and predefined. Also, a focus should be put on learning optimal precision weightings, depending on the current intention-context (action or observation), the representation in question, and maybe even depending on the interaction partner.

What should also be considered as a possible future direction, is the possibility to associate HPBU's actions with different intended effects than mere joint-movements. It would be interesting to associate effects in the world that do not influence the motor system directly. An example would be the distal action effect of pressing on a switch to turn on a light.

For a more fine-tuned, and possibly more appropriate belief coordination, which might be necessary for interaction with humans, representations at the coordination sequence level could also be learned. Generally, learning of new sensorimotor-part representations *during* social interaction would allow for very interesting investigations into human learning (e. g., language learning, or imitation learning), and negotiation.

Another angle for social interaction between social agents would be the variation of the agent's roles. In the present work, agents either have the role of *leader* or *follower*, while in both roles the agents were assumed to be collaborative. In future setups, the leader role could be configured (by means of different coordination sequences), to play a deceiving role in a deception game, explicitly trying to convince the follower agents of a false belief.

Tightly connected to such a setup is the idea of interaction-partner specific levels of trust. In the current model, if prior information about interaction partners are available, those beliefs will equally influence belief updating. In future interaction scenarios, the weight of prior information could, for example, be set to depend on the frequency of successful prior interactions with a specific interaction partner. This would make the mentioned deception game even more interesting.

As a final outlook, the possibility of future reimplementations are important and should strongly be encouraged. To my understanding, and as has previously been argued (Cooper and Guest, 2014), reimplementations could be regarded as replication. They make im-

plementation assumptions obvious and the basic assumptions are more rigorously tested. To that end, the source code for the HPBU core model as well as its specific instantiations (specifically configured for the presented evaluations), will be made public under open-source licencing*.

* Software repository: `https://purl.org/skahl/hpbu`

APPENDIX

A.1 FULL HIERARCHY OVERVIEW

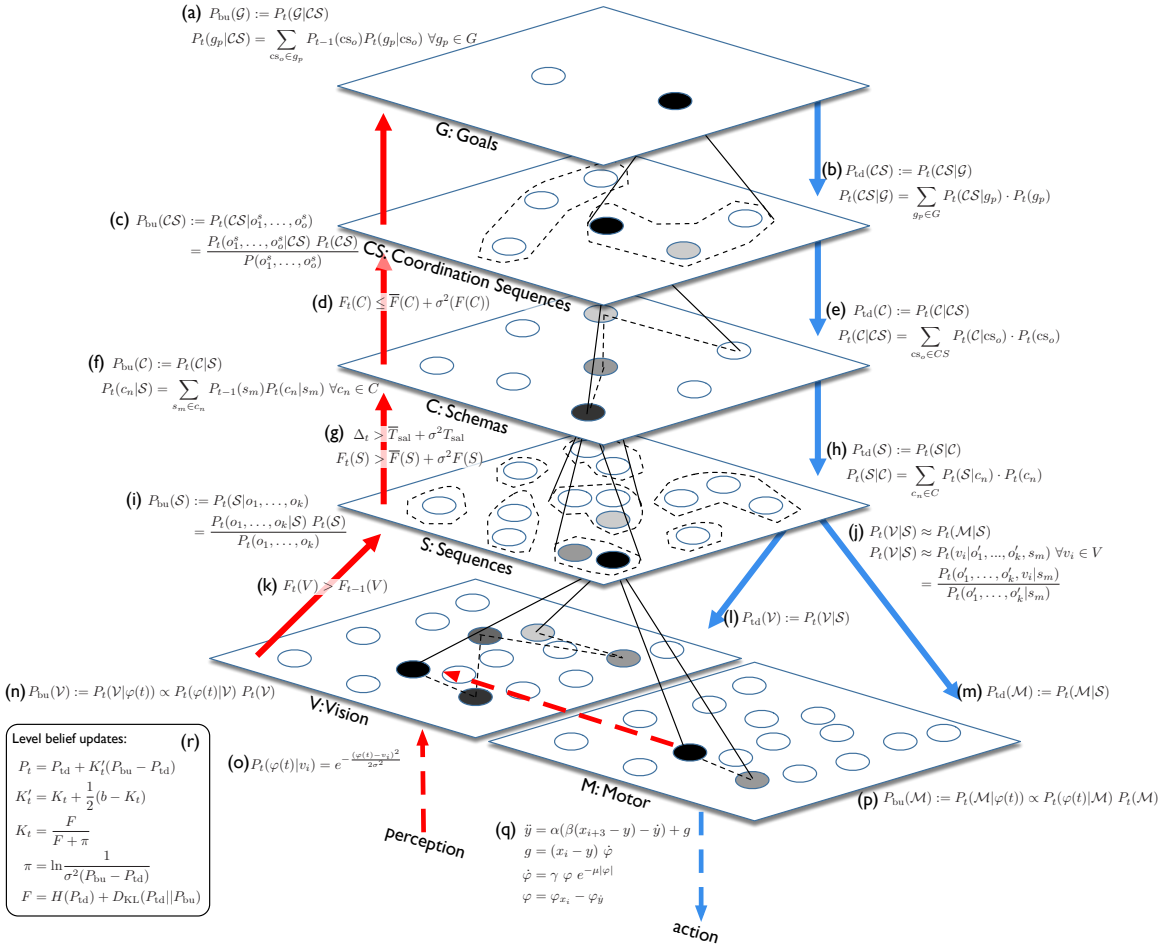


Figure A.1

Figure A.1 shows an overview over the whole HPBU hierarchy.

The two top-most levels make up the mentalizing part of HPBU, starting with level G (Goals), which shows: **(a)** bottom-up posterior of level G, see sec. 5.2.4; The next lower level CS (CoordinationSequences) shows: **(b)** top-down posterior of level CS, see sec. 5.2.4; **(c)** bottom-up posterior of level CS, see sec. 5.2.4.

Below that, the sensorimotor part of HPBU begings with level C (Schemas), which shows: **(d)** stable schema detection of level C, see par. 5.2.5; **(e)** top-down posterior of level C, see par. 4.2.1; **(f)** bottom-up posterior of level C, see par. 4.2.1.

Next, level S (Sequences) shows: **(g)** salient new sequence detection of level S, see par. 4.2.5; **(h)** top-down posterior of level S, see par. 4.2.1; **(i)** bottom-up posterior of level S, see par. 4.2.1; **(j)** likelihood of next steps of level V and M, see par. 4.2.1.

The bottom-most levels consist of levels V (Vision) and M (Motor-Control), which show: **(k)** salient movement detection of level V, see par. 4.2.3; **(l)** top-down posterior with (j) of level V; **(m)** top-down posterior with (j) of level M; **(n)** bottom-up posterior of level V, see par. 4.2.1; **(o)** likelihood of movement direction of level V, see par. 4.2.1; **(p)** bottom-up posterior of level M; **(q)** dampened spring system with goal-forcing, see par. 4.2.2.

The update of each level's beliefs are integrated using the follow belief update mechanism: **(r)** belief update, similar for all levels, see par. 4.1.4 and par. 4.2.4.

BIBLIOGRAPHY

Adams, R. A., S. Shipp, and K. J. Friston (Nov. 2012). "Predictions Not Commands: Active Inference in the Motor System". In: *Brain Struct. Funct.* 218.3, pp. 611–643. DOI: 10/f4wkqx (cit. on pp. 43, 80, 86).

Adams, R. A., K. E. Stephan, H. R. Brown, C. D. Frith, and K. J. Friston (2013). "The Computational Anatomy of Psychosis." In: *Front. Psychiatry* 4, p. 47. DOI: 10/gf7f92. pmid: 23750138 (cit. on pp. 32, 141).

Aglioti, S. M., P. Cesari, M. Romani, and C. Urgesi (Aug. 2008). "Action Anticipation and Motor Resonance in Elite Basketball Players". In: *Nat. Neurosci.* 11.9, pp. 1109–1116. DOI: 10/fqrn8w (cit. on p. 38).

Albrecht, D. W., I. Zukerman, and A. E. Nicholson (1998). "Bayesian Models for Keyhole Plan Recognition in an Adventure Game". In: *User Model. User-Adapt. Interact.* 8.1-2, pp. 5–47. DOI: 10/bh5fq9 (cit. on p. 59).

Alibali, M. W., D. C. Heath, and H. J. Myers (Feb. 2001). "Effects of Visibility between Speaker and Listener on Gesture Production: Some Gestures Are Meant to Be Seen". In: *J. Mem. Lang.* 44.2, pp. 169–188. DOI: 10/cjxsvs (cit. on pp. 21, 67).

Allwood, J. S., J. Nivre, and E. Ahlsén (1992). "On the Semantics and Pragmatics of Linguistic Feedback". In: p. 78. DOI: 10/d8wddm (cit. on p. 10).

Ambrosini, E., M. Ciavarro, G. Pelle, M. G. Perrucci, G. Galati, P. Fattori, C. Galletti, and G. Committeri (Dec. 2012). "Behavioral Investigation on the Frames of Reference Involved in Visuomotor Transformations during Peripheral Arm Reaching". In: *PLoS ONE* 7.12, e51856–8. DOI: 10/ggfj78 (cit. on pp. 75, 83).

Apperly, I. A. (Apr. 2008). "Beyond Simulation–Theory and Theory–Theory: Why Social Cognitive Neuroscience Should Use Its Own Concepts to Study Theory of Mind". In: *Cognition* 107.1, pp. 266–283. DOI: 10/bppvtx (cit. on p. 46).

Baker, C. L., R. Saxe, and J. B. Tenenbaum (2011). "Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution". In: *Proc. Annu. Meet. Cogn. Sci. Soc.* URL: https://cloudfront.escholarship.org/dist/prd/content/qt5rk7z59q/qt5rk7z59q.pdf (cit. on p. 59).

Barlow, H. B. (1961). "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication*. Ed. by W. A. Rosenblith. Cambridge, MA, pp. 217–234 (cit. on p. 41).

Baron-Cohen, S., A. M. Leslie, and U. Frith (Oct. 1985). "Does the Autistic Child Have a "Theory of Mind"?" In: *Cognition* 21.1, pp. 37–46. DOI: 10/c2nhzk. pmid: 2934210 (cit. on p. 44).

Barr, D. J. and B. Keysar (Feb. 2002). "Anchoring Comprehension in Linguistic Precedents". In: *J. Mem. Lang.* 46.2, pp. 391–418. DOI: 10/fc8v7w (cit. on p. 12).

Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston (Nov. 2012). "Canonical Microcircuits for Predictive Coding". In: *Neuron* 76.4, pp. 695–711. DOI: 10/f4gsgg (cit. on pp. 67, 72).

Bavelas, J., J. Gerwing, C. Sutton, and D. Prevost (Feb. 2008). "Gesturing on the Telephone: Independent Effects of Dialogue and Visibility". In: *J. Mem. Lang.* 58.2, pp. 495–520. DOI: 10/d4b42n (cit. on p. 21).

Becchio, C., A. Pierno, M. Mari, D. Lusher, and U. Castiello (Sept. 2007). "Motor Contagion from Gaze: The Case of Autism". In: *Brain* 130.9, pp. 2401–2411. DOI: 10/cqvpkn (cit. on p. 30).

Becchio, C., L. Sartori, and U. Castiello (June 2010). "Toward You: The Social Side of Actions". In: *Curr. Dir. Psychol. Sci.* 19.3, pp. 183–188. DOI: 10/br6dw4 (cit. on pp. 30, 46).

Benoit, R. G. and D. L. Schacter (Aug. 2015). "Specifying the Core Network Supporting Episodic Simulation and Episodic Memory by Activation Likelihood Estimation". In: *Neuropsychologia* 75, pp. 450–457. DOI: 10/f7p54p (cit. on pp. 48, 96).

Berniker, M. and K. Kording (Nov. 2008). "Estimating the Sources of Motor Errors for Adaptation and Generalization". In: *Nat. Neurosci.* 11.12, pp. 1454–1461. DOI: 10/fp2gqg (cit. on p. 56).

Bertenthal, B. I., M. R. Longo, and A. Kosobud (2006). "Imitative Response Tendencies Following Observation of Intransitive Actions." In: *J. Exp. Psychol. Hum. Percept. Perform.* 32.2, pp. 210–225. DOI: 10/fgtmnz (cit. on p. 26).

Bögels, S., D. J. Barr, S. Garrod, and K. Kessler (Aug. 2015). "Conversational Interaction in the Scanner: Mentalizing during Language Processing as Revealed by MEG". In: *Cereb. Cortex* 25.9, pp. 3219–3234. DOI: 10/f7rgf9 (cit. on p. 31).

Bohl, V. and N. Gangopadhyay (Dec. 2013). "Theory of Mind and the Unobservability of Other Minds". In: *Philos. Explor.* 17.2, pp. 203–222. DOI: 10/ggfj7z (cit. on p. 46).

Brandi, M.-L., D. Kaifel, D. Bolis, and L. Schilbach (Apr. 2019). "The Interactive Self – a Review on Simulating Social Interactions to Understand the Mechanisms of Social Agency". In: *-Com* 18.1, pp. 17–31. DOI: 10/ggfj7t (cit. on pp. 61, 149).

Brennan, S. E. (1990). *Seeking and Providing Evidence for Mutual Understanding*. URL: http://books.google.de/books?id=lbQgAQAAIAAJ&q=intitle:Seeking+and+providing+evidence+for+mutual+understanding&dq=intitle:Seeking+and+providing+evidence+for+mutual+understanding&hl=&cd=1&source=gbs_api (cit. on p. 12).

Brennan, S. E., A. Galati, and A. K. Kuhlen (2010). *Two Minds, One Dialog: Coordinating Speaking and Understanding*. 1st ed. Vol. 53. Elsevier Inc. DOI: 10.1016/S0079-7421(10)53008-1 (cit. on pp. 11, 12).

Brown, H., R. A. Adams, I. Parees, M. Edwards, and K. J. Friston (June 2013). "Active Inference, Sensory Attenuation and Illusions". In: *Cogn. Process.* 14.4, pp. 411–427. DOI: 10/f5cwh8 (cit. on pp. 32, 141).

Butz, M. V. (June 2016). "Toward a Unified Sub-Symbolic Computational Theory of Cognition". In: *Front. Psychol.* 7.500, pp. 611–19. DOI: 10/f8thgh (cit. on p. 57).

Butz, M. V., D. Bilkey, D. Humaidan, A. Knott, and S. Otte (Sept. 2018). "Learning, Planning, and Control in a Monolithic Neural Event Inference Architecture". In: *arXiv.org*. arXiv: 1809.07412v2 [cs.LG]. URL: http://arxiv.org/abs/1809.07412v2 (cit. on p. 58).

Calder, A. J., J. Keane, F. Manes, N. Antoun, and A. W. Young (Nov. 2000). "Impaired Recognition and Experience of Disgust Following Brain Injury." In: *Nat. Neurosci.* 3.11, pp. 1077–1078. DOI: 10/bjnz94. pmid: 11036262 (cit. on p. 23).

Catmur, C., R. B. Mars, M. F. Rushworth, and C. Heyes (Sept. 2011). "Making Mirrors: Premotor Cortex Stimulation Enhances Mirror and Counter-Mirror Motor Facilitation." In: *J. Cogn. Neurosci.* 23.9, pp. 2352–2362. DOI: 10/b45cqx. pmid: 20946056 (cit. on p. 25).

Cavallo, A., O. Lungu, C. Becchio, C. Ansuini, A. Rustichini, and L. Fadiga (Oct. 2015). "When Gaze Opens the Channel for Communication: Integrative Role of IFG and MPFC". In: *NeuroImage* 119.C, pp. 63–69. DOI: 10/f7rffq (cit. on p. 31).

Chambon, V. and P. Haggard (Dec. 2012). "Sense of Control Depends on Fluency of Action Selection, Not Motor Performance". In: *Cognition* 125.3, pp. 441–451. DOI: 10/gfs3cb (cit. on p. 54).

Chambon, V., N. Sidarus, and P. Haggard (May 2014). "From Action Intentions to Action Effects: How Does the Sense of Agency Come About?" In: *Front. Hum. Neurosci.* 8, pp. 1–9. DOI: 10/ggfj8b (cit. on pp. 54, 103, 121).

Charniak, E. and R. P. Goldman (Nov. 1993). "A Bayesian Model of Plan Recognition". In: *Proc. Annu. Meet. Cogn. Sci. Soc.* 64.1, pp. 53–79. DOI: 10/cjttdn (cit. on p. 58).

Chartrand, T. L. and J. A. Bargh (June 1999). "The Chameleon Effect: The Perception-Behavior Link and Social Interaction." In: *J. Pers. Soc. Psychol.* 76.6, pp. 893–910. ISSN: 0022-3514. DOI: 10/fbkx5m. pmid: 10402679 (cit. on pp. 8, 9).

Christianson, K., A. Hollingworth, J. F. Halliwell, and F. Ferreira (June 2001). "Thematic Roles Assigned along the Garden Path Linger". In: *Cognit. Psychol.* 42.4, pp. 368–407. DOI: 10/cv9jxk (cit. on p. 14).

Ciaramidaro, A., C. Becchio, L. Colle, B. G. Bara, and H. Walter (July 2014). "Do You Mean Me? Communicative Intentions Recruit the Mirror and the Mentalizing System". In: *Soc. Cogn. Affect. Neurosci.* 9.7, pp. 909–916. DOI: 10/f6bvpf (cit. on pp. 2, 31).

Cisek, P. (Sept. 2007). "Cortical Mechanisms of Action Selection: The Affordance Competition Hypothesis". In: *Philos. Trans. R. Soc. B Biol. Sci.* 362.1485, pp. 1585–1599. DOI: 10/ftxsfc (cit. on p. 49).

Clark, A. (2016). *Surfing Uncertainty*. Prediction, Action, and the Embodied Mind. Oxford University Press, USA. ISBN: 0-19-021701-4. URL: http://books.google.de/books?id=Yoh2CgAAQBAJ&printsec=frontcover&dq=intitle:Surfing+Uncertainty&hl=&cd=1&source=gbs_api (cit. on pp. 2, 41, 43, 56, 66).

Clark, H. H. and S. E. Brennan (1991). "Grounding in Communication". In: *Perspect. Socially Shar. Cogn.* DOI: 10/c9wt3w (cit. on pp. 10, 12, 13).

Clark, H. H. (1996). "Using Language". In: *Comput. Linguist.* 23, p. 452. ISSN: 00318094. DOI: 10/ggfj8j. pmid: 2561775 (cit. on pp. 1, 4, 8, 9, 11, 18).

Clark, H. H. and M. A. Krych (Jan. 2004). "Speaking While Monitoring Addressees for Understanding". In: *J. Mem. Lang.* 50.1, pp. 62–81. DOI: 10/fbqprb (cit. on p. 13).

Clark, H. H. and E. F. Schaefer (Feb. 1989). "Contributing to Discourse". In: *Cogn. Sci.* 13.2, pp. 259–294. DOI: 10/btp9wg (cit. on pp. 1, 11, 13).

Colonius, H. and A. Diederich (July 2004). "Multisensory Interaction in Saccadic Reaction Time: A Time-Window-of-Integration Model". In: *J. Cogn. Neurosci.* 16.6, pp. 1000–1009. DOI: 10/dx9nkt (cit. on p. 52).

Condon, W. S. and W. D. Ogston (Oct. 1966). "Sound Film Analysis of Normal and Pathological Behavior Patterns." In: *J. Nerv. Ment. Dis.* 143.4, pp. 338–347. DOI: 10/ddqg3r. pmid: 5958766 (cit. on p. 9).

Cooper, R. P. and O. Guest (Mar. 2014). "Implementations Are Not Specifications: Specification, Replication and Experimentation in Computational Cognitive Modeling". In: *Cogn. Syst. Res.* 27.C, pp. 42–49. DOI: 10/ggfj8n (cit. on pp. 65, 133, 150).

Cox, R. T. (Jan. 1946). "Probability, Frequency and Reasonable Expectation". In: *Am. J. Phys.* 14.1, pp. 1–13. DOI: 10/dm54sw (cit. on p. 36).

Cunningham, D. W., V. A. Billock, and B. H. Tsou (Nov. 2001). "Sensorimotor Adaptation to Violations of Temporal Contiguity." In: *Psychol. Sci.* 12.6, pp. 532–535. DOI: 10/fjqphm. pmid: 11760144 (cit. on p. 51).

Darwiche, A. (Apr. 2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press. ISBN: 0-521-88438-1. URL: http://books.google.de/books?id=HsGrXdZMXg4C&printsec=frontcover&dq=intitle:Modeling+and+Reasoning+with+Bayesian+Networks&hl=&cd=1&source=gbs_api (cit. on p. 77).

Dayan, P., G. E. Hinton, R. M. Neal, and R. S. Zemel (Sept. 1995). "The Helmholtz Machine". In: *Neural Comput.* 7.5, pp. 889–904. DOI: 10/cqbn3w. pmid: [objectObject] (cit. on pp. 37, 88).

De la Rosa, S., S. Streuber, M. Giese, H. H. Bülthoff, and C. Curio (Jan. 2014). "Putting Actions in Context: Visual Action Adaptation

Aftereffects Are Modulated by Social Contexts". In: *PLoS ONE* 9.1, e86502–10. DOI: 10/ggfj8s (cit. on p. 26).

De Bruin, L. and D. Strijbos (Nov. 2015). "Direct Social Perception, Mindreading and Bayesian Predictive Coding." In: *Conscious. Cogn.* 36, pp. 565–570. DOI: 10/ggfj8t. pmid: 25959592 (cit. on p. 46).

De Cruys, S. V., K. Evers, R. Van der Hallen, L. Van Eylen, B. Boets, L. de-Wit, and J. Wagemans (Oct. 2014). "Precise Minds in Uncertain Worlds: Predictive Coding in Autism". In: *Psychol. Rev.* 121.4, pp. 649–675. DOI: 10/f6nc7k (cit. on p. 141).

Dean, T. and K. Kanazawa (1989). "A Model for Reasoning about Persistence and Causation". In: *Comput. Intell. Neurosci.* 5, pp. 142–150. DOI: 10/fw6xhz (cit. on p. 59).

Decety, J. and J. A. Sommerville (Dec. 2003). "Shared Representations between Self and Other: A Social Cognitive Neuroscience View". In: *Trends Cogn. Sci.* 7.12, pp. 527–533. DOI: 10/dnfdn7 (cit. on pp. 2, 22).

Dennett, D. C. (1978). "Beliefs about Beliefs". In: *Behav. Brain Sci.* 1.4, p. 568. DOI: 10/cfzp8g (cit. on p. 43).

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press. ISBN: 978-0-262-54053-7. URL: http://books.google.de/books?id=Qbvkja-J9iQC&printsec=frontcover&dq=intitle:The+Intentional+Stance&hl=&cd=1&source=gbs_api (cit. on p. 52).

Devaine, M., G. Hollard, and J. Daunizeau (Dec. 2014). "The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn?" In: *PLoS Comput. Biol.* 10.12, e1003992–14. DOI: 10/ggfj8q (cit. on p. 45).

Di Pellegrino, G., L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti (1992). "Understanding Motor Events: A Neurophysiological Study." In: *Exp. Brain Res.* 91.1, pp. 176–180. DOI: 10/btr2pk. pmid: 1301372 (cit. on p. 23).

Doiega, K. (2018). "Commentary: M-Autonomy." In: *Front. Psychol.* 9, p. 680. DOI: 10/ggf756. pmid: 29899713 (cit. on p. 139).

Donnarumma, F., H. Dindo, P. Iodice, and G. Pezzulo (Mar. 2017). "You Cannot Speak and Listen at the Same Time: A Probabilistic Model of Turn-Taking". In: *Biol. Cybern.* 111.2, pp. 1–19. DOI: 10/f945dg (cit. on p. 60).

Dumas, G., M. Chavez, J. Nadel, and J. Martinerie (May 2012a). "Anatomical Connectivity Influences Both Intra- and Inter-Brain Synchronizations". In: *PLoS ONE* 7.5, e36414–11. DOI: 10/f3x74d (cit. on pp. 62, 146).

Dumas, G., J. Martinerie, R. Soussignan, and J. Nadel (2012b). "Does the Brain Know Who Is at the Origin of What in an Imitative Interaction?" In: *Front. Hum. Neurosci.* 6, pp. 1–12. DOI: 10/ggfj8d (cit. on p. 53).

Dumas, G., J. Nadel, R. Soussignan, J. Martinerie, and L. Garnero (Aug. 2010). "Inter-Brain Synchronization during Social Interaction". In: *PLoS ONE* 5.8, e12166–10. DOI: 10/ccb5mv (cit. on p. 61).

Dunbar, R. I. M. (Jan. 1998). "The Social Brain Hypothesis". In: *Evol. Anthropol. Issues News Rev.* 6.5, pp. 178–190. DOI: 10/cdgqn8 (cit. on p. 24).

Engel, K. C., M. Flanders, and J. F. Soechting (July 2002). "Oculocentric Frames of Reference for Limb Movement." In: *Arch. Ital. Biol.* 140.3, pp. 211–219. pmid: 12173524. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=12173524&retmode=ref&cmd=prlinks (cit. on pp. 75, 83).

Eshghi, A., C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver (2015). "Feedback in Conversation as Incremental Semantic Update". In: *aclweb.org*. URL: https://www.aclweb.org/anthology/W15-0130 (cit. on p. 16).

Feldman, A. G. and M. F. Levin (Dec. 1995). "The Origin and Use of Positional Frames of Reference in Motor Control". In: *Behav. Brain Sci.* 18.4, pp. 723–744. DOI: 10/brftsd (cit. on p. 80).

Ferreira, F. and J. Stacey (2000). "The Misinterpretation of Passive Sentences". In: *Citeseer*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.7728&rep=rep1&type=pdf (cit. on p. 14).

Ferreira, F., K. G. D. Bailey, and V. Ferraro (Feb. 2002). "Good-Enough Representations in Language Comprehension". In: *Curr. Dir. Psychol. Sci.* 11.1, pp. 11–15. DOI: 10/fhm2mr (cit. on pp. 1, 14).

Fogassi, L. and G. Luppino (Dec. 2005). "Motor Functions of the Parietal Lobe". In: *Curr. Opin. Neurobiol.* 15.6, pp. 626–631. DOI: 10/dxzh33 (cit. on p. 66).

Fotopoulou, A. and M. Tsakiris (Apr. 2017). "Mentalizing Homeostasis: The Social Origins of Interoceptive Inference". In: *Neuropsychoanalysis* 19.1, pp. 3–28. DOI: 10/ggfj8c (cit. on p. 61).

Frey, B. J. and D. Dueck (Feb. 2007). "Clustering by Passing Messages between Data Points." In: *Science* 315.5814, pp. 972–976. DOI: 10/c6vwpc. pmid: 17218491 (cit. on pp. 90, 91).

Friston, K. J. and C. D. Frith (1995). "Schizophrenia: A Disconnection Syndrome?" In: *Clin. Neurosci. N. Y. N* 3.2, pp. 89–97. pmid: 7583624. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=7583624&retmode=ref&cmd=prlinks (cit. on p. 141).

Friston, K. J. (2011). "What Is Optimal about Motor Control?" In: *Neuron* 72.3, pp. 488–498. DOI: 10/b3fk4b. pmid: 22078508 (cit. on pp. 2, 57, 79, 81, 121).

Friston, K. J. (Sept. 2013). "Life as We Know It." In: *J. R. Soc. Interface R. Soc.* 10.86, pp. 20130475–20130475. DOI: 10/x22. pmid: [objectObject] (cit. on pp. 2, 4, 39, 66).

Friston, K. J., J. Daunizeau, J. Kilner, and S. J. Kiebel (Feb. 2010). "Action and Behavior: A Free-Energy Formulation". In: *Biol. Cybern.* 102.3, pp. 227–260. DOI: 10/c8zhrp (cit. on pp. 2, 39, 43).

Friston, K. J. and C. D. Frith (Nov. 2015a). "A Duet for One." In: *Conscious. Cogn.* 36, pp. 390–405. DOI: 10/gf97fh. pmid: 25563935 (cit. on p. 49).

Friston, K. J. and C. D. Frith (July 2015b). "Active Inference, Communication and Hermeneutics." In: *CORTEX* 68, pp. 129–143. DOI: 10/f7jc6p. pmid: 25957007 (cit. on pp. 49, 62).

Friston, K. J. and S. Kiebel (May 2009). "Predictive Coding under the Free-Energy Principle". In: 364.1521, pp. 1211–1221. DOI: 10/fcswfc (cit. on pp. 2, 41, 66).

Friston, K. J., T. Parr, and B. de Vries (2017a). "The Graphical Brain: Belief Propagation and Active Inference." In: *Netw. Neurosci. Camb. Mass* 1.4, pp. 381–414. DOI: 10/gfrbcv. pmid: 29417960 (cit. on p. 70).

Friston, K. J., R. Rosch, T. Parr, C. Price, and H. Bowman (June 2017b). "Deep Temporal Models and Active Inference". In: *Neurosci. Biobehav. Rev.* 77, pp. 388–402. DOI: 10/gbgmrg (cit. on pp. 135, 146, 149).

Friston, K. J., T. Shiner, T. FitzGerald, J. M. Galea, R. Adams, H. Brown, R. J. Dolan, R. Moran, K. E. Stephan, and S. Bestmann (Jan. 2012). "Dopamine, Affordance and Active Inference". In: *PLoS Comput. Biol.* 8.1, e1002327–20. DOI: 10/fxmcrc (cit. on p. 139).

Frith, C. D., Blakemore, and D. M. Wolpert (Dec. 2000). "Abnormalities in the Awareness and Control of Action". In: *Philos. Trans. R. Soc. B Biol. Sci.* 355.1404, pp. 1771–1788. DOI: 10/b4jkkn (cit. on p. 51).

Frith, C. D. (Aug. 2012). "The Role of Metacognition in Human Social Interactions." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 367.1599, pp. 2213–2223. DOI: 10/gddc2b. pmid: [objectObject] (cit. on pp. 45, 134).

Frith, C. D. and U. Frith (Nov. 2008). "Implicit and Explicit Processes in Social Cognition". In: *Neuron* 60.3, pp. 503–510. DOI: 10/d9xrnh (cit. on p. 45).

Froese, T. and D. a Leavens (2014). "The Direct Perception Hypothesis: Perceiving the Intention of Another's Action Hinders Its Precise Imitation." In: *Front. Psychol.* 5 (February), pp. 65–15. ISSN: 16641078. DOI: 10/f6tr9t. pmid: [objectObject] (cit. on p. 31).

Gallagher, S. (Oct. 2005). *How the Body Shapes the Mind*. Clarendon Press. ISBN: 0-19-162257-5. URL: http://books.google.de/books?id=1FuOy1jPK3UC&printsec=frontcover&dq=intitle:How+the+body+shapes+the+mind&hl=&cd=1&source=gbs_api (cit. on pp. 2, 22).

Gallagher, S. (June 2008). "Direct Perception in the Intersubjective Context." In: *Conscious. Cogn.* 17.2, pp. 535–543. DOI: 10/fckj3g. pmid: 18442924 (cit. on pp. 31, 44, 46).

Gallagher, S. (Nov. 2015). "The New Hybrids: Continuing Debates on Social Perception". In: *Conscious. Cogn.* 36.C, pp. 452–465. DOI: 10/ggfj8f (cit. on p. 46).

Gallese, V., L. Fadiga, L. Fogassi, and G. Rizzolatti (Apr. 1996). "Action Recognition in the Premotor Cortex." In: *Brain* 119 (Pt 2), pp. 593–

609. pmid: 8800951. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=8800951&retmode=ref&cmd=prlinks (cit. on pp. 2, 23).

Gallese, V. and A. Goldman (Dec. 1998). "Mirror Neurons and the Simulation Theory of Mind-Reading." In: *Trends Cogn. Sci.* 2.12, pp. 493–501. DOI: 10/c3kjb6. pmid: 21227300 (cit. on p. 44).

Gallese, V., C. Keysers, and G. Rizzolatti (Sept. 2004). "A Unifying View of the Basis of Social Cognition." In: *Trends Cogn. Sci.* 8.9, pp. 396–403. DOI: 10/fpm4f6. pmid: 15350240 (cit. on p. 24).

Gangopadhyay, N. and L. Schilbach (July 2012). "Seeing Minds: A Neurophilosophical Investigation of the Role of Perception-Action Coupling in Social Perception." In: *Soc. Neurosci.* 7.4, pp. 410–423. DOI: 10/fp6zc9. pmid: 22059802 (cit. on p. 29).

Garrod, S. and M. J. Pickering (Jan. 2004). "Why Is Conversation so Easy?" In: *Trends Cogn. Sci.* 8.1, pp. 8–11. DOI: 10/d4zrv7. pmid: 14697397 (cit. on p. 10).

Gentile, G., M. Björnsdotter, V. I. Petkova, Z. Abdulkarim, and H. H. Ehrsson (Apr. 2015). "Patterns of Neural Activity in the Human Ventral Premotor Cortex Reflect a Whole-Body Multisensory Percept". In: *NeuroImage* 109.C, pp. 328–340. DOI: 10/f3nm8m (cit. on p. 66).

Georgiou, I., C. Becchio, S. Glover, and U. Castiello (Mar. 2007). "Different Action Patterns for Cooperative and Competitive Behaviour". In: *Cognition* 102.3, pp. 415–433. DOI: 10/fxfjhg (cit. on p. 26).

Gigerenzer, G., P. M. Todd, A. B. C. R. Group, and n. others (1999). "Simple Heuristics That Make Us Smart". In: (cit. on p. 15).

Giles, H. and N. Coupland (1991). *Language: Contexts and Consequences*. Belmond, CA: Wadsworth Publishing. URL: https://psycnet.apa.org/record/1992-97484-000 (cit. on p. 9).

Goldin-Meadow, S. (June 2006). "Talking and Thinking with Our Hands". In: *Psychol. Sci.* 15.1, pp. 34–39. DOI: 10/fvxxmj (cit. on p. 20).

Goldin-Meadow, S. and S. L. Beilock (Dec. 2010). "Actions Influence on Thought: The Case of Gesture". In: *Perspect. Psychol. Sci.* 5.6, pp. 664–674. DOI: 10/fgk7xd (cit. on pp. 2, 22, 67).

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative Adversarial Nets". In: pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets (cit. on p. 112).

Gopnik, A. and H. M. Wellman (Mar. 1992). "Why the Child's Theory of Mind Really Is a Theory". In: *Mind Lang.* 7.1-2, pp. 145–171. DOI: 10/fng26v (cit. on p. 44).

Gordon, R. M. (June 1986). "Folk Psychology as Simulation". In: *Mind Lang.* 1.2, pp. 158–171. DOI: 10/bwmv6r (cit. on p. 44).

Graziano, M. S., D. F. Cooke, and C. S. Taylor (Dec. 2000). "Coding the Location of the Arm by Sight." In: *Science* 290.5497, pp. 1782–1786. DOI: 10/bdbq77. pmid: 11099420 (cit. on p. 66).

Gregory, R. L. (July 1980). "Perceptions as Hypotheses." In: *Philos. Trans. R. Soc. B Biol. Sci.* 290.1038, pp. 181–197. DOI: 10/cgdwx9. pmid: 6106237 (cit. on p. 36).

Grice, H. P. (1975). "Logic and Conversation". In: *Syntax and Semantics 3*. Elsevier, pp. 41–58. ISBN: 0-12-785423-1 (cit. on pp. 11, 133).

Griffiths, T. L., N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum (Aug. 2010). "Probabilistic Models of Cognition: Exploring Representations and Inductive Biases." In: *Trends Cogn. Sci.* 14.8, pp. 357–364. DOI: 10/csgrqm. pmid: 20576465 (cit. on p. 55).

Grossman, E. D. and R. Blake (2001). "Brain Activity Evoked by Inverted and Imagined Biological Motion." In: *Vision Res.* 41.10-11, pp. 1475–1482. DOI: 10/c9h7ps. pmid: 11322987 (cit. on p. 23).

Gumbsch, C., S. Otte, and M. V. Butz (2017). "A Computational Model for the Dynamical Learning of Event Taxonomies". In: *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. London, UK, pp. 452–457. URL: https://pdfs.semanticscholar.org/bbab/29f9342f5aaf4da94936d5da2275b963aa2b.pdf (cit. on p. 88).

Haering, C. and A. Kiesel (Mar. 2015). "Was It Me When It Happened Too Early? Experience of Delayed Effects Shapes Sense of Agency". In: *Cognition* 136.C, pp. 38–42. DOI: 10/f62779 (cit. on p. 51).

Haggard, P. and S. Clark (Dec. 2003). "Intentional Action: Conscious Experience and Neural Prediction". In: *Conscious. Cogn.* 12.4, pp. 695–707. DOI: 10/fcs6x3 (cit. on p. 53).

Haggard, P., S. Clark, and J. Kalogeras (Apr. 2002). "Voluntary Action and Conscious Awareness." In: *Nat. Neurosci.* 5.4, pp. 382–385. DOI: 10/cj73gg. pmid: [objectObject] (cit. on p. 53).

Han, K. and M. Veloso (2000). "Automated Robot Behavior Recognition". In: *Robotics Research*. London: Springer, London, pp. 249–256. ISBN: 978-1-4471-1254-9. DOI: 10.1007/978-1-4471-0765-1_30 (cit. on p. 59).

Hassabis, D. and E. A. Maguire (July 2007). "Deconstructing Episodic Memory with Construction". In: *Trends Cogn. Sci.* 11.7, pp. 299–306. DOI: 10/db6mf6 (cit. on pp. 48, 50, 96).

Healey, P. G. T., G. J. Mills, A. Eshghi, and C. Howes (Apr. 2018). "Running Repairs: Coordinating Meaning in Dialogue". In: *Top. Cogn. Sci.* 10.2, pp. 367–388. DOI: 10/gd8tpw (cit. on pp. 1, 16, 17, 133).

Heider, F. and M. Simmel (1944). "An Experimental Study of Apparent Behavior". In: *Am. J. Psychol.* 57.2, pp. 243–259. DOI: 10/ftcck7. JSTOR: 1416950?origin=crossref (cit. on pp. 27, 43).

Helmholtz, H. von (1867). *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss (cit. on p. 36).

Heyes, C. (June 2001). "Causes and Consequences of Imitation." In: *Trends Cogn. Sci.* 5.6, pp. 253–261. DOI: 10/dbcwd8. pmid: 11390296 (cit. on p. 25).

Heyes, C. (Nov. 2009). "Where Do Mirror Neurons Come From?" In: *Neurosci. Biobehav. Rev.*, pp. 1–9. DOI: 10/b3qmx3 (cit. on p. 25).

Hillock-Dunn, A. and M. T. Wallace (Aug. 2012). "Developmental Changes in the Multisensory Temporal Binding Window Persist into Adolescence". In: *Dev. Sci.* 15.5, pp. 688–696. DOI: 10/f38wjw (cit. on p. 52).

Hinton, G. E. and R. S. Zemel (1994). "Autoencoders, Minimum Description Length and Helmholtz Free Energy". In: *Advances in Neural Information Processing Systems*. Ed. by J. Cowan, G. Tesauro, and A. J. San Mateo, CA, pp. 1–8 (cit. on p. 39).

Hochreiter, S. and J. Schmidhuber (Nov. 1997). "Long Short-Term Memory." In: *Neural Comput.* 9.8, pp. 1735–1780. DOI: 10/bxd65w. pmid: 9377276 (cit. on p. 57).

Hoffmann, H., P. Pastor, D.-H. Park, and S. Schaal (2009). "Biologically-Inspired Dynamical Systems for Movement Generation: Automatic Real-Time Goal Adaptation and Obstacle Avoidance". In: *2009 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 2587–2592. ISBN: 978-1-4244-2788-8. DOI: 10/dxj9qk (cit. on p. 80).

Hommel, B., J. Müsseler, G. Aschersleben, and W. Prinz (Oct. 2001). "The Theory of Event Coding (TEC): A Framework for Perception and Action Planning." In: *Behav. Brain Sci.* 24.5, 849-78- discussion 878–937. ISSN: 0140-525X. DOI: 10/brbjv2. pmid: [objectObject] (cit. on p. 26).

Hosoya, T., S. A. Baccus, and M. Meister (July 2005). "Dynamic Predictive Coding by the Retina". In: *Nature* 436.7047, pp. 71–77. DOI: 10/c4zwss (cit. on p. 40).

Hurley, S. (Feb. 2008). "The Shared Circuits Model (SCM): How Control, Mirroring, and Simulation Can Enable Imitation, Deliberation, and Mindreading." In: *Behav. Brain Sci.* 31.1, 1-22- discussion 22–58. DOI: 10/fh36zd. pmid: 18394222 (cit. on p. 46).

Hutchison, W. D., K. D. Davis, A. M. Lozano, R. R. Tasker, and J. O. Dostrovsky (May 1999). "Pain-Related Neurons in the Human Cingulate Cortex." In: *Nat. Neurosci.* 2.5, pp. 403–405. DOI: 10/c37rqp. pmid: 10321241 (cit. on p. 23).

Ijspeert, A. J., J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal (Feb. 2013). "Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors". In: *MIT Press* 25.2, pp. 328–373. DOI: 10/f4hm9v (cit. on p. 80).

Jacob, P. and M. Jeannerod (Jan. 2005). "The Motor Theory of Social Cognition: A Critique". In: *Trends Cogn. Sci.* 9.1, pp. 21–25. DOI: 10/bmthhs (cit. on p. 27).

James, W. (1890). *The Principles of Psychology*. In Two Volumes. Holt. ISBN: 779595833. URL: http://books.google.de/books?id=VR41xgEACAAJ&dq=intitle:The+principles+of+psychology&hl=&cd=8&source=gbs_api (cit. on p. 42).

Jeannerod, M. (July 2001). "Neural Simulation of Action: A Unifying Mechanism for Motor Cognition". In: *NeuroImage* 14.1, S103–S109. DOI: 10/c23gsd (cit. on p. 44).

Jordan, J. S. (Dec. 1998). "Recasting Dewey's Critique of the Reflex-Arc Concept via a Theory of Anticipatory Consciousness: Implications for Theories of Perception". In: *Proc. Annu. Meet. Cogn. Sci. Soc.* 16.3, pp. 165–187. DOI: 10/fp6vkh (cit. on p. 88).

Jordan, M. I. and D. E. Rumelhart (Sept. 1992). "Forward Models: Supervised Learning with a Distal Teacher". In: *Proc. Annu. Meet. Cogn. Sci. Soc.* 16.3, pp. 307–354. DOI: 10/cdk3vh (cit. on p. 37).

Kahl, S. and S. Kopp (2015). "Towards a Model of the Interplay of Mentalizing and Mirroring in Embodied Communication". In: *EuroAsianPacific Joint Conference on Cognitive Science* (cit. on pp. 60, 63).

Kahl, S. and S. Kopp (2018). "A Predictive Processing Model of Perception and Action for Self-Other Distinction". In: *Frontiers in Psychology* (cit. on pp. 50, 56, 73, 79, 103, 114, 118).

Kawato, M. and H. Gomi (1992). "A Computational Model of Four Regions of the Cerebellum Based on Feedback-Error Learning." In: *Biol. Cybern.* 68.2, pp. 95–103. DOI: 10/dxm4z8. pmid: 1486143 (cit. on p. 37).

Kawato, M., H. Hayakawa, and T. Inui (1993). "A Forward-Inverse Optics Model of Reciprocal Connections between Visual Cortical Areas". In: *Taylor Francis*. URL: https://www.tandfonline.com/doi/abs/10.1088/0954-898X_4_4_001 (cit. on pp. 37, 88).

Kawato, M. (Dec. 1999). "Internal Models for Motor Control and Trajectory Planning". In: *Curr. Opin. Neurobiol.* 9.6, pp. 718–727. DOI: 10/fhswb8 (cit. on p. 56).

Keller, P. E., G. Knoblich, and B. H. Repp (Mar. 2007). "Pianists Duet Better When They Play with Themselves: On the Possible Role of Action Simulation in Synchronization". In: *Conscious. Cogn.* 16.1, pp. 102–111. DOI: 10/cmwkkf (cit. on p. 38).

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press (cit. on p. 56).

Kelso, J. A. S., G. Dumas, and E. Tognoli (Jan. 2013). "Outline of a General Theory of Behavior and Brain Coordination". In: *Neural Netw.* 37, pp. 120–131. DOI: 10/f2dkh2 (cit. on p. 62).

Kendon, A. (1973). "The Role of Visible Behavior in the Organization of Social Interaction". In: *Social Communication and Movement*. Ed. by M. Cranach and I. Vine. London: Wiley, pp. 29–74. URL: https://ci.nii.ac.jp/naid/10009701793/ (cit. on p. 9).

Keysar, B., S. Lin, and D. J. Barr (Aug. 2003). "Limits on Theory of Mind Use in Adults". In: *Cognition* 89.1, pp. 25–41. DOI: 10/cjtdfn (cit. on p. 45).

Keysers, C., B. Wicker, V. Gazzola, J.-L. Anton, L. Fogassi, and V. Gallese (Apr. 2004). "A Touching Sight: SII/PV Activation during

the Observation and Experience of Touch." In: *Neuron* 42.2, pp. 335–346. DOI: 10/fwqzp8. pmid: 15091347 (cit. on pp. 2, 23).

Khalsa, S. S. and R. C. Lapidus (July 2016). "Can Interoception Improve the Pragmatic Search for Biomarkers in Psychiatry?" In: *Front. Psychiatry* 7 (Suppl), p. 633. DOI: 10/ggfj7r (cit. on p. 61).

Kiebel, S. J., K. Von Kriegstein, and J. Daunizeau (2009). "Recognizing Sequences of Sequences". In: *PLoS Comput. Biol.* DOI: 10/bn5d6t (cit. on p. 67).

Kilner, J. M., Y. Paulignan, and S. J. Blakemore (Mar. 2003). "An Interference Effect of Observed Biological Movement on Action". In: *Curr. Biol.* 13.6, pp. 522–525. DOI: 10/bnw2ns (cit. on p. 45).

Kingma, D. P. and J. Ba (Dec. 2017). "Adam: A Method for Stochastic Optimization". In: *arXiv.org*. arXiv: 1412.6980v9 [cs.LG]. URL: http://arxiv.org/abs/1412.6980v9 (cit. on p. 112).

Knill, D. C. and A. Pouget (Dec. 2004). "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation". In: *Trends Neurosci.* 27.12, pp. 712–719. DOI: 10/fgfb9j (cit. on pp. 37, 55).

Knoblich, G., R. Flach, and 2001 (May 2016). "Predicting the Effects of Actions: Interactions of Perception and Action". In: *Psychol. Sci.* 12.6, pp. 467–472. DOI: 10/dfh64b (cit. on p. 38).

Kohonen, T. (1983). "Self-Organizing Feature Maps". In: *Self-Organization and Associative Memory*. Berlin, Heidelberg (cit. on p. 40).

Kok, P., D. Rahnev, J. F. M. Jehee, H. C. Lau, and F. P. de Lange (Sept. 2012). "Attention Reverses the Effect of Prediction in Silencing Sensory Signals." In: *Cereb. Cortex* 22.9, pp. 2197–2206. DOI: 10/ccgpk7. pmid: 22047964 (cit. on p. 42).

Konvalinka, I., P. Vuust, A. Roepstorff, and C. D. Frith (Nov. 2010). "Follow You, Follow Me: Continuous Mutual Prediction and Adaptation in Joint Tapping". In: 63.11, pp. 2220–2230. DOI: 10/c8bpxr (cit. on p. 17).

Kopp, S. (June 2010). "Social Resonance and Embodied Coordination in Face-to-Face Conversation with Artificial Interlocutors". In: *Speech Commun.* 52.6, pp. 587–597. DOI: 10/b6wfs4 (cit. on pp. 8, 13).

Krauss, R. M., Y. Chen, and R. F. Gottesmann (2001). "Lexical Gestures and Lexical Access: A Process Model". In: *Language and Gesture*. Ed. by D. McNeill. books.google.com, pp. 261–283 (cit. on pp. 2, 20).

Kullback, S. (1959). *Statistics and Information Theory*. New York: Wiley (cit. on p. 37).

Kuramoto, Y. (1975). "Self-Entrainment of a Population of Coupled Non-Linear Oscillators". In: *Int. Symp. Math. Probl.* DOI: 10/c99jz4 (cit. on p. 62).

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum (2015). "Human-Level Concept Learning through Probabilistic Program Induction". In: *Science* 350.6266, pp. 1332–1338. DOI: 10/f73zfg. pmid: 26659050 (cit. on p. 68).

Lakin, J. L., V. E. Jefferis, C. M. Cheng, and T. L. Chartrand (2003). "The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry". In: *J. Nonverbal Behav.* 27.3, pp. 145–162. DOI: 10/dhq2ct (cit. on p. 9).

Laughlin, S. B. and R. C. Hardie (1978). "Common Strategies for Light Adaptation in the Peripheral Visual Systems of Fly and Dragonfly". In: *J. Comp. Physiol. A* 128.4, pp. 319–340. DOI: 10/c37wgb (cit. on p. 41).

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (Nov. 1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proc. IEEE* 86.11, pp. 2278–2324. DOI: 10/d89c25 (cit. on p. 67).

Maguire, E. A. and C. J. Mummery (Jan. 1999). "Differential Modulation of a Common Memory Retrieval Network Revealed by Positron Emission Tomography". In: *Hippocampus* 9.1, pp. 54–61. DOI: 10/b29dhf (cit. on pp. 48, 96).

Malle, B. F. (2001). "Folk Explanations of Intentional Action". In: *Intentions and Intentionality: Foundations of Social Cognition* (cit. on p. 10).

Manning, C. D., P. Raghavan, and H. Schütze (July 2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN: 1-139-47210-0. URL: http://books.google.de/books?id=t1PoSh4uwVcC&pg=PP1&dq=intitle:Introduction+to+Information+Retrieval&hl=&cd=1&source=gbs_api (cit. on p. 73).

Marcus, G. F. (Aug. 2010). "Neither Size Fits All: Comment on McClelland et al. and Griffiths et Al." In: *Trends Cogn. Sci.* 14.8, pp. 346–347. DOI: 10/b2c8zw (cit. on p. 55).

Mareschal, I., Y. Otsuka, and C. W. G. Clifford (Oct. 2014). "A Generalized Tendency toward Direct Gaze with Uncertainty". In: *J. Vis.* 14.12, pp. 27–27. DOI: 10/f6qqd3 (cit. on p. 30).

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. URL: http://93.174.95.29/_ads/FD35491AB493C4FAC30E01B03225AE51 (cit. on pp. 3, 7, 132).

Mateos, D. M., L. E. Riveaud, and P. W. Lamberti (Aug. 2017). "Detecting Dynamical Changes in Time Series by Using the Jensen Shannon Divergence". In: *Chaos Interdiscip. J. Nonlinear Sci.* 27.8, pp. 083118–14. DOI: 10/gbvj7d (cit. on p. 83).

Mayberry, R. I., J. Jaques, and G. DeDe (1998). "What Stuttering Reveals about the Development of the Gesture-Speech Relationship." In: *New Dir. Child Dev.* 1998.79, pp. 77–87. DOI: 10/ggdcnr. pmid: 9507705 (cit. on p. 20).

McClelland, J. L., M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith (Aug. 2010). "Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Cognition". In: *Trends Cogn. Sci.* 14.8, pp. 348–356. DOI: 10/dncj2p (cit. on p. 55).

McNeill, D. (Sept. 2008). *Gesture and Thought*. University of Chicago Press. ISBN: 978-0-226-51464-2. URL: http://books.google.de/books?id=N0SmyU4TKRwC&printsec=frontcover&dq=intitle:Gesture+and+thought&hl=&cd=1&source=gbs_api (cit. on pp. 21, 67).

McNeill, D. and S. D. Duncan (2000). "Growth Points in Thinking-for-Speaking". In: ed. by D. McNeill, pp. 141–161 (cit. on pp. 2, 20, 21).

Metzinger, T. (2015). "M-Autonomy". In: *J. Conscious. Stud.* URL: https://www.ingentaconnect.com/content/imp/jcs/2015/00000022/F0020011/art00013 (cit. on p. 139).

Minsky, M. (June 1974). "A Framework for Representing Knowledge". In: URL: https://dspace.mit.edu/handle/1721.1/6089 (cit. on p. 47).

Moore, J. W. and P. C. Fletcher (Mar. 2012). "Sense of Agency in Health and Disease: A Review of Cue Integration Approaches". In: *Conscious. Cogn.* 21.1, pp. 59–68. DOI: 10/bvtb52 (cit. on pp. 53, 120).

Moore, J. W., D. M. Wegner, and P. Haggard (Dec. 2009). "Modulating the Sense of Agency with External Cues". In: *Conscious. Cogn.* 18.4, pp. 1056–1064. DOI: 10/fk38dq (cit. on p. 53).

Mukamel, R., A. D. Ekstrom, J. Kaplan, M. Iacoboni, and I. Fried (Apr. 2010). "Single-Neuron Responses in Humans during Execution and Observation of Actions". In: *CURBIO* 20.8, pp. 750–756. DOI: 10/dfp4j6 (cit. on p. 23).

Mumford, D. (1992). "On the Computational Architecture of the Neocortex. II. The Role of Cortico-Cortical Loops." In: *Biol. Cybern.* 66.3, pp. 241–251. DOI: 10/fc636h. pmid: 1540675 (cit. on pp. 40, 41).

Myllyneva, A. and J. K. Hietanen (Jan. 2015). "There Is More to Eye Contact than Meets the Eye". In: *Cognition* 134, pp. 100–109. DOI: 10/f6wr47 (cit. on pp. 10, 30).

Nahab, F. B., P. Kundu, C. Gallea, J. Kakareka, R. Pursley, T. Pohida, N. Miletta, J. Friedman, and M. Hallett (Dec. 2010). "The Neural Processes Underlying Self-Agency". In: *Cereb. Cortex* 21.1, pp. 48–55. DOI: 10/ff5tpr (cit. on p. 53).

Necker, L. A. (June 1832). "LXI. Observations on Some Remarkable Optical Phænomena Seen in Switzerland; and on an Optical Phænomenon Which Occurs on Viewing a Figure of a Crystal or Geometrical Solid". In: *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 1.5, pp. 329–337. DOI: 10/c44wtf (cit. on p. 36).

Otte, S., T. Schmitt, K. J. Friston, and M. V. Butz (Sept. 2017). "Inferring Adaptive Goal-Directed Behavior within Recurrent Neural Networks". In: *Artificial Neural Networks and Machine Learning – ICANN 2017*. Cham: Springer, Cham, pp. 227–235. ISBN: 978-3-319-68599-1. DOI: 10.1007/978-3-319-68600-4_27 (cit. on p. 58).

Parr, T. and K. J. Friston (Sept. 2019). "Generalised Free Energy and Active Inference". In: *Biol. Cybern.*, pp. 1–19. DOI: 10/gf85p4 (cit. on p. 71).

Penrose, L. S. and R. Penrose (Apr. 1958). "IMPOSSIBLE OBJECTS: A SPECIAL TYPE OF VISUAL ILLUSION". In: *Br. J. Psychol.* 49.1, pp. 31–33. DOI: 10/cdc4v3 (cit. on p. 37).

Pezzulo, G. and P. Cisek (June 2016). "Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition." In: *Trends Cogn. Sci.* 20.6, pp. 414–424. DOI: 10/f8pb23. pmid: 27118642 (cit. on p. 49).

Pezzulo, G. and H. Dindo (May 2011). "What Should I Do next? Using Shared Representations to Solve Interaction Problems". In: *Exp. Brain Res.* 211.3-4, pp. 613–630. DOI: 10/d734b6 (cit. on p. 60).

Pezzulo, G., F. Donnarumma, and H. Dindo (2013). "Human Sensorimotor Communication: A Theory of Signaling in Online Social Interactions." In: *PLoS ONE* 8.11, e79876. DOI: 10/ggfj8r. pmid: 24278201 (cit. on pp. 2, 4, 18, 104).

Pfister, R., T. Melcher, A. Kiesel, P. Dechent, and O. Gruber (Feb. 2014). "Neural Correlates of Ideomotor Effect Anticipations". In: *NSC* 259.C, pp. 164–171. DOI: 10/ggfj7s (cit. on p. 43).

Pickering, M. J. and S. Frisson (2001). "Processing Ambiguous Verbs: Evidence from Eye Movements." In: *J. Exp. Psychol. Learn. Mem. Cogn.* 27.2, p. 556. DOI: 10/cnmf2p (cit. on p. 14).

Pickering, M. J. and S. Garrod (Apr. 2004). "Toward a Mechanistic Psychology of Dialogue". In: *Behav. Brain Sci.* 27.2, 169-90- discussion 190–226. DOI: 10.1017/S0140525X04000056. pmid: [objectObject] (cit. on pp. 8, 12).

Pipereit, K., O. Bock, and J.-L. Vercher (Mar. 2006). "The Contribution of Proprioceptive Feedback to Sensorimotor Adaptation". In: *Exp. Brain Res.* 174.1, pp. 45–52. DOI: 10/fwmkxg (cit. on p. 66).

Pöppel, J. and S. Kopp (2019). *Satisficing Mentalizing: Bayesian Models of Theory of Mind Reasoning in Scenarios with Different Uncertainties*. URL: http://arxiv.org/abs/1909.10419 (cit. on p. 59).

Posner, M. I. (1980). "Orienting of Attention". In: *Q. J. Exp. Psychol.* 32.1, pp. 3–25. DOI: 10/bh4h54 (cit. on p. 42).

Premack, D. and G. Woodruff (Dec. 1978). "Does the Chimpanzee Have a Theory of Mind?" In: *Behav. Brain Sci.* 1.4, pp. 515–526. DOI: 10/ddvt4n (cit. on pp. 16, 43).

Press, C., G. Bird, E. Walsh, and C. Heyes (June 2008). "Automatic Imitation of Intransitive Actions". In: *BRAIN Cogn.* 67.1, pp. 44–50. DOI: 10/bjw2h9 (cit. on p. 26).

Press, C., P. Kok, and D. Yon (Aug. 2019). "The Perceptual Prediction Paradox". In: *psyarxiv.com*. DOI: 10/ggfj72 (cit. on p. 42).

Prinz, W. (1990). "A Common Coding Approach to Perception and Action". In: *Relationships between Perception and Action*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, pp. 167–201. ISBN: 978-3-642-75350-3. DOI: 10.1007/978-3-642-75348-0_7 (cit. on pp. 2, 26, 42, 80).

Prinz, W. (1997). "Perception and Action Planning". In: *Eur. J. Cogn. Psychol.* DOI: 10/dgfm52 (cit. on p. 26).

Przyrembel, M., J. Smallwood, M. Pauen, and T. Singer (June 2012). "Illuminating the Dark Matter of Social Neuroscience: Considering the Problem of Social Interaction from Philosophical, Psychological, and Neuroscientific Perspectives". In: *Front. Hum. Neurosci.* 6, pp. 1–15. DOI: 10/gfgrjq (cit. on pp. 2, 29, 146).

Pynadath, D. V. and S. Marsella (2005). "PsychSim: Modeling Theory of Mind with Decision-Theoretic Agents". In: *ccs.neu.edu*. URL: http://www.ccs.neu.edu/home/marsella/publications/pdf/PynMarsIJCAI05.pdf (cit. on p. 59).

Rabinowitz, N. C., F. Perbet, H. F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick (Feb. 2018). "Machine Theory of Mind". In: *arXiv.org*. arXiv: 1802.07740v2 [cs.AI]. URL: http://arxiv.org/abs/1802.07740v2 (cit. on p. 59).

Raichle, M. E., A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman (Jan. 2001). "A Default Mode of Brain Function." In: *Proc. Natl. Acad. Sci.* 98.2, pp. 676–682. DOI: 10/djhbks. pmid: 11209064 (cit. on p. 28).

Raichle, M. E. and D. A. Gusnard (Aug. 2002). "Appraising the Brain's Energy Budget." In: *Proc. Natl. Acad. Sci.* 99.16, pp. 10237–10239. DOI: 10/fr2gxx. pmid: [objectObject] (cit. on pp. 38, 66).

Raichle, M. E. and A. Z. Snyder (Oct. 2007). "A Default Mode of Brain Function: A Brief History of an Evolving Idea". In: *NeuroImage* 37.4, pp. 1083–1090. DOI: 10/c2gcw6 (cit. on p. 28).

Rao, A. S. and M. P. Georgeff (1995). "BDI Agents: From Theory to Practice." In: *Proc. First Int. Conf. Multiagent Syst.* URL: https://www.aaai.org/Papers/ICMAS/1995/ICMAS95-042.pdf (cit. on pp. 16, 58).

Rao, R. P. and D. H. Ballard (Jan. 1999). "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects." In: *Nat. Neurosci.* 2.1, pp. 79–87. DOI: 10/drddxm. pmid: 10195184 (cit. on pp. 39, 40).

Real, E., S. Moore, A. Selle, S. Saxena, and Y. L. Suematsu (2017). "Large-Scale Evolution of Image Classifiers". In: *arXiv.org*. scholar: E3EF07AC-0D7F-42CE-A19F-F3BC82EA87F0. URL: http://scholar.google.comjavascript:void(0) (cit. on p. 112).

Reddy, M. (1979). "The Conduit Metaphor - A Case of Frame Conflict in Our Language about Language". In: *Metaphor and Thought*. Ed. by A. Ortony. Cambridge, pp. 284–310. ISBN: 0-521-29626-9. URL: http://scholar.google.comjavascript:void(0) (cit. on p. 8).

Reddy, V. (Sept. 2003). "On Being the Object of Attention: Implications for Self–Other Consciousness". In: *Trends Cogn. Sci.* 7.9, pp. 397–402. DOI: 10/d294kw (cit. on p. 30).

Rizzolatti, G. and M. A. Arbib (May 1998). "Language within Our Grasp." In: *Trends Neurosci.* 21.5, pp. 188–194. DOI: 10/fbx8qz. pmid: 9610880 (cit. on p. 24).

Rohde, M. and M. O. Ernst (Apr. 2016). "Time, Agency, and Sensory Feedback Delays during Action". In: *Curr. Opin. Behav. Sci.* 8, pp. 193–199. DOI: 10/ggfj8h (cit. on p. 51).

Roth, N. A. (Mar. 2012). "A Note on the Gesture of Writing by Vilém Flusser and the Gesture of Writing". In: *New Writ.* 9.1, pp. 24–41. DOI: 10/ggfj76 (cit. on p. 21).

Rumelhart, D. E. (1975). "Notes on Schema for Stories". In: *Representation and Understanding*. Ed. by D. G. Bobrow and A. Collins, pp. 211–236 (cit. on p. 47).

Russo, G. S. and C. J. Bruce (Aug. 1996). "Neurons in the Supplementary Eye Field of Rhesus Monkeys Code Visual Targets and Saccadic Eye Movements in an Oculocentric Coordinate System". In: *J. Neurophysiol.* 76.2, pp. 825–848. DOI: 10/ggfj79 (cit. on pp. 75, 83).

Sachs, J. S. (1967). "Recopition Memory for Syntactic and Semantic Aspects of Connected Discourse". In: *Percept. Psychophys.* 2.9, pp. 437–442. DOI: 10/fwb274 (cit. on p. 14).

Sacks, H., E. A. Schegloff, and G. Jefferson (1978). *A Simplest Systematics for the Organization of Turn Taking for Conversation.* J. Schenkein. New York: Academic Press. URL: http://scholar.google.com/scholar?q=related:uDms73pGXNAJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5&as_ylo=1978&as_yhi=1978 (cit. on p. 16).

Sadeghipour, A. and S. Kopp (Nov. 2010). "Embodied Gesture Processing: Motor-Based Integration of Perception and Action in Social Artificial Agents". In: *Cogn. Comput.* 3.3, pp. 419–435. DOI: 10/b8hnvg (cit. on p. 60).

Salvador, S., P. Chan, and 2004 (Jan. 2007). "Toward Accurate Dynamic Time Warping in Linear Time and Space". In: *Intell. Data Anal.* 11.5, pp. 561–580. DOI: 10/gfn732 (cit. on p. 83).

Sanford, A. and P. Sturt (Sept. 2002). "Depth of Processing in Language Comprehension: Not Noticing the Evidence." In: *Trends Cogn. Sci.* 6.9, p. 382. DOI: 10/b3gzgz. pmid: [objectObject] (cit. on p. 14).

Schegloff, E. A. (June 1987). "Analyzing Single Episodes of Interaction: An Exercise in Conversation Analysis". In: *Soc. Psychol. Q.* 50.2, p. 101. DOI: 10/b3zfjh. JSTOR: 2786745?origin=crossref (cit. on p. 16).

Schegloff, E. A. (1995). "Discourse as an Interactional Achievement III: The Omnirelevance of Action." In: *Res. Lang. Soc. Interact.* (28(3)), pp. 185–211. DOI: 10/fwjgx6 (cit. on p. 16).

Schilbach, L., S. B. Eickhoff, A. Rotarska-Jagiela, G. R. Fink, and K. Vogeley (June 2008). "Minds at Rest? Social Cognition as the Default Mode of Cognizing and Its Putative Relationship to the Default System of the Brain". In: *Conscious. Cogn.* 17.2, pp. 457–467. DOI: 10/bfqv2p (cit. on pp. 2, 29, 30).

Schilbach, L., B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, and K. Vogeley (Aug. 2013). "Toward a Second-Person Neuroscience." In: *Behav. Brain Sci.* 36.4, pp. 393–414. ISSN: 1469-1825. DOI: 10/f45g7b. pmid: 23883742 (cit. on pp. 30, 46, 141).

Schilling, M. and H. Cruse (2012). "Whats next: Recruitment of a Grounded Predictive Body Model for Planning a Robots Actions". In: *Front. Psychol.* 3, pp. 1–19. DOI: 10/ggfj7v (cit. on p. 57).

Schuwerk, T., B. Langguth, and M. Sommer (Nov. 2014). "Modulating Functional and Dysfunctional Mentalizing by Transcranial Magnetic Stimulation". In: *Front. Psychol.* 5 (November), pp. 1–9. DOI: 10/f6q7nv (cit. on pp. 2, 28).

Sebanz, N., H. Bekkering, and G. Knoblich (Feb. 2006). "Joint Action: Bodies and Minds Moving Together". In: *Trends Cogn. Sci.* 10.2, pp. 70–76. DOI: 10/fs6rd9 (cit. on p. 45).

Shafir, E. and A. Tversky (Oct. 1992). "Thinking through Uncertainty: Nonconsequential Reasoning and Choice." In: *Cognit. Psychol.* 24.4, pp. 449–474. DOI: 10/d6thrq. pmid: 1473331 (cit. on p. 55).

Sherwell, C., M. Garrido, and R. Cunnington (Dec. 2016). "Timing in Predictive Coding: The Roles of Task Relevance and Global Probability." In: *J. Cogn. Neurosci.* 29.5, pp. 1–13. DOI: 10/gf2bwc. pmid: 27991186 (cit. on p. 51).

Shockley, K., M.-V. Santana, and C. A. Fowler (Apr. 2003). "Mutual Interpersonal Postural Constraints Are Involved in Cooperative Conversation." In: *J. Exp. Psychol. Hum. Percept. Perform.* 29.2, pp. 326–332. DOI: 10/dgjdcj. pmid: [objectObject] (cit. on p. 9).

Shultz, T. R. and K. Cloghesy (1981). "Development of Recursive Awareness of Intention." In: *Dev. Psychol.* 17.4, pp. 465–471. DOI: 10/d4rtcc (cit. on p. 45).

Sidarus, N., V. Chambon, and P. Haggard (Dec. 2013). "Priming of Actions Increases Sense of Control over Unexpected Outcomes". In: *Conscious. Cogn.* 22.4, pp. 1403–1411. DOI: 10/f5h7z6 (cit. on p. 52).

Simon, H. A. (Feb. 1955). "A Behavioral Model of Rational Choice". In: *Q. J. Econ.* 69.1, pp. 99–118. DOI: 10.2307/1884852 (cit. on p. 14).

Sokoloff, L., R. Mangold, R. L. Wechsler, C. Kenney, and S. S. Kety (July 1955). "The Effect of Mental Arithmetic on Cerebral Circulation and Metabolism." In: *J. Clin. Invest.* 34 (7, Part 1), pp. 1101–1108. DOI: 10/bbpwhg. pmid: 14392225 (cit. on pp. 38, 66).

Sommer, M. A. and R. H. Wurtz (July 2008). "Brain Circuits for the Internal Monitoring of Movements*". In: *Annu. Rev. Neurosci.* 31.1, pp. 317–338. DOI: 10/dxqf63 (cit. on p. 86).

Spivey, M. (June 2008). *The Continuity of Mind*. Oxford University Press, USA. ISBN: 0-19-803815-1. URL: http://books.google.de/books?id=FLZII0uI0kgC&pg=PA9&dq=intitle:The+Continuity+of+Mind&hl=&cd=1&source=gbs_api (cit. on p. 88).

Stone, V. E., S. Baron-Cohen, and R. T. Knight (Sept. 1998). "Frontal Lobe Contributions to Theory of Mind." In: *J. Cogn. Neurosci.* 10.5, pp. 640–656. DOI: 10/dh4znz. pmid: 9802997 (cit. on p. 44).

Street, R. L. (Dec. 1984). "Speech Convergence and Speech Evaluation in Fact-Finding Interviews". In: *Hum. Commun. Res.* 11.2, pp. 139–169. DOI: 10/bps399 (cit. on p. 9).

Streuber, S., G. Knoblich, N. Sebanz, H. H. Bülthoff, and S. de la Rosa (Aug. 2011). "The Effect of Social Context on the Use of Visual Information". In: *Exp. Brain Res.* 214.2, pp. 273–284. DOI: 10/cn3tdq (cit. on p. 26).

Swets, B. and F. Ferreira (2002). "How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums". In: 46.1, pp. 57–84 (cit. on pp. 10, 15).

Swets, B., M. E. Jacovina, and R. J. Gerrig (Jan. 2013). "Effects of Conversational Pressures on Speech Planning". In: *Discourse Process.* 50.1, pp. 23–51. DOI: 10/ggfj8m (cit. on p. 15).

Synofzik, M., G. Vosgerau, and A. Newen (Mar. 2008). "Beyond the Comparator Model: A Multifactorial Two-Step Account of Agency." In: *Conscious. Cogn.* 17.1, pp. 219–239. DOI: 10/bqnzk5. pmid: 17482480 (cit. on p. 53).

Synofzik, M., G. Vosgerau, and M. Voss (2013). "The Experience of Agency: An Interplay between Prediction and Postdiction." In: *Front. Psychol.* 4, p. 127. DOI: 10/gbfpnf. pmid: 23508565 (cit. on pp. 53, 120).

Tatler, B. W., M. M. Hayhoe, M. F. Land, and D. H. Ballard (May 2011). "Eye Guidance in Natural Vision: Reinterpreting Salience". In: *J. Vis.* 11.5, pp. 5–5. DOI: 10/d3cdkt (cit. on p. 42).

Teufel, C., P. C. Fletcher, and G. Davis (Aug. 2010). "Seeing Other Minds: Attributed Mental States Influence Perception". In: *Trends Cogn. Sci.* 14.8, pp. 376–382. DOI: 10/bzf6bz (cit. on p. 31).

Tickle-Degnen, L. and R. Rosenthal (Oct. 1990). "The Nature of Rapport and Its Nonverbal Correlates". In: *Psychol. Inq.* 1.4, pp. 285–293. DOI: 10/dh7pmt (cit. on p. 8).

Todorov, E. and M. I. Jordan (Nov. 2002). "Optimal Feedback Control as a Theory of Motor Coordination". In: *Nat. Neurosci.* 5.11, pp. 1226–1235. DOI: 10/drpq85 (cit. on pp. 56, 81).

Tomasello, M. (Aug. 2008). *Origins of Human Communication*. MIT Press. ISBN: 0-262-26120-0. URL: http://books.google.de/books?id=T3bqzIe3mAEC&printsec=frontcover&dq=intitle:Origins+of+Human+Communication&hl=&cd=1&source=gbs_api (cit. on pp. 2, 13, 19, 67).

Traum, D. R. and J. Allen (1992). "A "Speech Acts" Approach to Grounding in Conversation". In: *International Conference on Spoken Language Processing*. URL: https://www.isca-speech.org/archive/archive_papers/icslp_1992/i92_0137.pdf (cit. on pp. 1, 10).

Treue, S. and J. C. Martinez Trujillo (June 1999). "Feature-Based Attention Influences Motion Processing Gain in Macaque Visual Cortex." In: *Nature* 399.6736, pp. 575–579. DOI: 10/c6xvwp. pmid: 10376597 (cit. on p. 138).

Tulving, E. (Jan. 1985). "Memory and Consciousness." In: *Can. Psychol. Can.* 26.1, pp. 1–12. DOI: 10/cnz69p (cit. on p. 47).

Tversky, A. and D. Kahneman (Sept. 1973). "Availability: A Heuristic for Judging Frequency and Probability". In: *Cognit. Psychol.* 5.2, pp. 207–232. DOI: 10/c47mgk (cit. on p. 138).

Van der Weiden, A., M. Prikken, and N. E. M. van Haren (Oct. 2015). "Self-Other Integration and Distinction in Schizophrenia: A Theoretical Analysis and a Review of the Evidence." In: *Neurosci. Biobehav. Rev.* 57, pp. 220–237. DOI: 10/f7xjfw. pmid: 26365106 (cit. on pp. 32, 50–52, 141).

Van Overwalle, F. (Mar. 2009). "Social Cognition and the Brain: A Meta-Analysis". In: *Hum. Brain Mapp.* 30.3, pp. 829–858. DOI: 10/bqd7c7 (cit. on pp. 2, 24, 25, 28, 29, 96).

Van Overwalle, F. (Jan. 2011). "A Dissociation between Social Mentalizing and General Reasoning". In: *NeuroImage* 54.2, pp. 1589–1599. DOI: 10/cscgwb (cit. on pp. 48, 50).

Van Baaren, R. B., R. W. Holland, K. Kawakami, and A. van Knippenberg (May 2016). "Mimicry and Prosocial Behavior". In: *Psychol. Sci.* 15.1, pp. 71–74. DOI: 10/c7kthh (cit. on p. 9).

Vesper, C. and M. J. Richardson (May 2014). "Strategic Communication and Behavioral Coupling in Asymmetric Joint Action". In: *Exp. Brain Res.* 232.9, pp. 2945–2956. DOI: 10/f6c8wv (cit. on pp. 2, 17).

Vesper, C., L. Schmitz, L. Safra, N. Sebanz, and G. Knoblich (Aug. 2016). "The Role of Shared Visual Information for Joint Action Coordination". In: *Cognition* 153.C, pp. 118–123. DOI: 10/f8t26b (cit. on p. 18).

Vilares, I. and K. Kording (Apr. 2011). "Bayesian Models: The Structure of the World, Uncertainty, Behavior, and the Brain". In: *Ann. N. Y. Acad. Sci.* 1224.1, pp. 22–39. DOI: 10/bh2wrk (cit. on p. 55).

Von der Lühe, T., V. Manera, I. Barisic, C. Becchio, K. Vogeley, and L. Schilbach (May 2016). "Interpersonal Predictive Coding, Not Action Perception, Is Impaired in Autism." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371.1693, p. 20150373. DOI: 10/ggdbsg. pmid: 27069050 (cit. on p. 141).

Wang, Y. and A. F. d. C. Hamilton (2013). "Understanding the Role of the 'self' in the Social Priming of Mimicry." In: *PLoS ONE* 8.4, e60249. DOI: 10/f4rz93. pmid: 23565208 (cit. on p. 9).

Wason, P. C. (July 1960). "On the Failure to Eliminate Hypotheses in a Conceptual Task". In: *Q. J. Exp. Psychol.* 12.3, pp. 129–140. DOI: 10/cms3r4 (cit. on p. 138).

Wegner, D. M. and T. Wheatley (July 1999). "Apparent Mental Causation. Sources of the Experience of Will." In: *Am. Psychol.* 54.7, pp. 480–492. DOI: 10/d56643. pmid: 10424155 (cit. on p. 52).

Weiss, C., A. Herwig, and S. Schütz-Bosbach (July 2011). "The Self in Social Interactions: Sensory Attenuation of Auditory Action Effects Is Stronger in Interactions with Others". In: *PLoS ONE* 6.7, e22723–3. DOI: 10/fwx4jb (cit. on p. 51).

Willems, R. M. and J. C. Francken (2012). "Embodied Cognition: Taking the next Step." In: *Front. Psychol.* 3 (December), p. 582. ISSN: 1664-1078. DOI: 10/ggfj8k. pmid: [objectObject] (cit. on p. 26).

Wilson, A. D. and S. Golonka (2013). "Embodied Cognition Is Not What You Think It Is." In: *Front. Psychol.* 4 (February), p. 58. ISSN: 1664-1078. DOI: 10/gfj8pf. pmid: 23408669 (cit. on p. 15).

Wilson, M. (Dec. 2002). "Six Views of Embodied Cognition." In: *Psychon. Bull. Rev.* 9.4, pp. 625–636. ISSN: 1069-9384. DOI: 10/fhrj23. pmid: [objectObject] (cit. on pp. 2, 21, 26, 66).

Wimmer, H. and J. Perner (Jan. 1983). "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." In: *Cognition* 13.1, pp. 103–128. ISSN: 00100277. DOI: 10/cbrpxb. pmid: [objectObject] (cit. on p. 44).

Wise, S. P., D. Boussaoud, P. B. Johnson, and R. Caminiti (1997). "Premotor and Parietal Cortex: Corticocortical Connectivity and Combinatorial Computations." In: *Annu. Rev. Neurosci.* 20.1, pp. 25–42. DOI: 10/bd5wgm. pmid: 9056706 (cit. on p. 66).

Wohlschläger, A., M. Gattis, and H. Bekkering (Feb. 2003). "Action Generation and Action Perception in Imitation: An Instance of the Ideomotor Principle". In: *Philos. Trans. R. Soc. B Biol. Sci.* 358.1431, pp. 501–515. DOI: 10/cwdx8w (cit. on p. 42).

Wolpe, N., J. W. Moore, C. L. Rae, T. Rittman, E. Altena, P. Haggard, and J. B. Rowe (Jan. 2014). "The Medial Frontal-Prefrontal Network for Altered Awareness and Control of Action in Corticobasal Syndrome". In: *Brain* 137.1, pp. 208–220. DOI: 10/f5qqsb (cit. on pp. 53, 120).

Wolpert, D. M., K. Doya, and M. Kawato (Mar. 2003). "A Unifying Computational Framework for Motor Control and Social Interaction". In: *Philos. Trans. R. Soc. B Biol. Sci.* 358.1431, pp. 593–602. DOI: 10/fwpndf (cit. on pp. 57, 146).

Wolpert, D. M., Z. Ghahramani, and M. I. Jordan (Sept. 1995). "An Internal Model for Sensorimotor Integration." In: *Science* 269.5232, pp. 1880–1882. DOI: 10/c2kdcc. pmid: 7569931 (cit. on pp. 37, 86).

Wolpert, D. M. and M. Kawato (1998). "Multiple Paired Forward and Inverse Models for Motor Control". In: *Neural Netw.* 11.7-8, pp. 1317–1329. DOI: 10/bksjsp. pmid: 12662752 (cit. on p. 63).

Wykowska, A., E. Wiese, A. Prosser, and H. J. Müller (Apr. 2014). "Beliefs about the Minds of Others Influence How We Process Sensory

Information". In: *PLoS ONE* 9.4, e94339–11. DOI: 10/f3sc54 (cit. on p. 31).

Zacks, J. M. and B. Tversky (Jan. 2001). "Event Structure in Perception and Conception." In: *Psychol. Bull.* 127.1, pp. 3–21. DOI: 10/bt5skb. pmid: 11271755 (cit. on p. 47).

Zacks, J. M., N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds (2007). "Event Perception: A Mind-Brain Perspective." In: *Psychol. Bull.* 133.2, pp. 273–293. DOI: 10/d34fh2 (cit. on pp. 47, 82, 88).

DECLARATION

Hiermit erkläre ich, dass ich diese Dissertation konform zu § 8 Abs. 1 lit g der Rahmenpromotionsordnung der Universität Bielefeld vom 15. Juni 2010 angefertigt habe, d. h.

- mir ist die geltende Promotionsordnung der Technischen Fakultät der Universität Bielefeld vom 1. März 2011 bekannt;
- ich habe die Dissertation selbst angefertigt, keine Textabschnitte von Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen in meiner Arbeit angegeben;
- Dritte haben weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Vermittlungstätigkeiten oder für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- diese Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht; und
- die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung wurde von mir bei keiner anderen Hochschule als Dissertation eingereicht.

Bielefeld, 24.02.2020

Sebastian Kahl