UNIVERSITÄTSKLINIKUM HAMBURG-EPPENDORF

Institut für Systemische Neurowissenschaften

Direktor: Prof. Dr. med. Christian Büchel

Do multidimensional representations of personality support learning?

Dissertation

zur Erlangung des Doktorgrades Dr. rer. biol. hum.

an der Medizinischen Fakultät der Universität Hamburg.

vorgelegt von:

Koen Frolichs

Aus Heerlen, Niederlande

Hamburg 2024

Angenommen von der
Medizinischen Fakultät der Universität Hamburg am: **29.05.2024**

Veröffentlicht mit Genehmigung der
Medizinischen Fakultät der Universität Hamburg.

Prüfungsausschuss, der/die Vorsitzende: **Prof. Dr. Christoph Korn**

Prüfungsausschuss, zweite/r Gutachter/in: **Prof. Dr. Nicolas Schuck**

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor Christoph Korn. In Christoph's lab, I could flourish because there was always time for discussions about science (among other things), endless encouragement, unbounded optimism, and unwavering support. A meeting with Christoph would never last the intended time because our shared enthusiasm made time go by too fast. This effect magnified whenever Gabriela Rosenblau joined our discussions. I feel incredibly lucky to have also had Gabriela as an unofficial supervisor. Gabriela helped me tremendously with writing, clarifying concepts, and general science, and occasionally offered much needed emotional support.

Two people who have been with me almost from the beginning of my PhD journey are Mana Ehlers and Lisa Doppelhofer. Navigating science, the PhD, and life was so much smoother with you two there. Our Sunday bouldering sessions at FLASHH were a highlight of every week. Thank you so much for your friendship, patience, and proofreading. To Lisa, who so graciously offered me her (purely metaphorical) shoulder to cry on when the last months got tough, I will be forever grateful.

To all the members of the DNHI-lab (new and old) who made lunch fun, lab-meetings long, and science scintillating I express my thanks (Christoph, Lisa, Benjamin, Sihui, Sergej, Ismail, and Xiong). Thanks to Ben, Lisa, and Sihui who, at different times, tolerated sharing an office with me.

The group of friends I made during my master. Who know firsthand the difficulties of science and who were always there to give advice and lend an ear: Jan, Till, Alina and Julio. Meeting with you all was so refreshing and fun, I cannot wait for the next time we are all together again.

The people at the ISN who made me feel at home right away. Everyone who attended methods meetings and of course the members of Sgt. Pepper's Lonely Math club (specific thanks to Uli and Tobias). Thanks for making my stay in Hamburg unforgettable. To the friends I made along the way: Anna, Alice, and Giulia I am happy you are only a phone call away!

Finally, my family, whose unwavering support made me who I am and gave me the strength to push through when the going got tough.

Abstract

In this thesis, I would like to explore how humans learn about others' personalities. First, from a behavioral perspective, I want to understand the strategies underlying this learning, i.e., find out what information people use and how they apply it. To do this, I use reinforcement learning models that are commonly used to explain learning across domains. I do not use these models on their own but add information structures that humans might use when learning about others. These models reveal that humans flexibly use complex knowledge structures when learning about others' personalities. Based on these results, I next extend my focus from behavior to the brain. In recent years, developments in analysis techniques of functional brain data have allowed for queries into activity patterns rather than plain activations. Such analyses revealed multidimensional structures when learning about others' personalities. I use these techniques to find out whether the cortex also codes for even more complex information structures that I found in the behavioral study. In general, I find evidence that the brain represents these complex knowledge structures when learning about strangers' personalities. Finally, in the most recent work, I focus on grid-cells, which represent a specific way for representing and coding information. First established in navigational space, grid-like coding patterns have also been found for conceptual spaces in two dimensions. I investigate a personality plane based on two trait words that form the axes of this 'trait space' but with the used analysis methods, I find no evidence for grid-like coding for this trait space.

Zusammenfassung (summary in German)

In dieser Dissertation hab ich untersucht, wie Menschen über die Persönlichkeit anderer lernen. Zuerst möchte ich aus einer Verhaltensperspektive die Strategien verstehen, die diesem Lernen zugrunde liegen, d. h. ich möchte herausfinden, welche Informationen Menschen nutzen und wie sie diese anwenden. Dazu verwende ich "Reinforcement Learning" Modelle, die häufig verwendet werden, um Lernen in verschiedensten Bereichen zu erklären. Diese Modelle verwende ich nicht isoliert, sondern ergänze sie um Informationsstrukturen, die Menschen beim Lernen über andere möglicherweise nutzen. Durch diese Modelle wird sichtbar, dass Menschen flexibel komplexe Informationsstrukturen gebrauchen, wenn sie über die Persönlichkeit anderer Personen lernen. Auf der Grundlage dieser Ergebnisse erweitere ich meinen Fokus vom Verhalten auf das Gehirn. In den letzten Jahren haben die Entwicklungen der Analysetechniken für funktionelle Hirndaten die Möglichkeit eröffnet, Aktivitätsmuster und nicht nur einfache Aktivierungen zu untersuchen. Solche Analysen enthüllten mehrdimensionale Strukturen beim Lernen über die Persönlichkeiten anderer. Ich nutze diese Techniken, um herauszufinden, ob das Gehirn auch für noch komplexere Informationsstrukturen kodiert, wie solche die ich in der Verhaltensstudie gefunden habe. Grundsätzlich gibt es Hinweise darauf, dass das Gehirn diese komplexen Wissensstrukturen repräsentiert, wenn man über die Persönlichkeit Fremder lernt. Schließlich konzentriere ich mich in meiner letzten Studie auf sogenannte Grid-Cells, die eine besondere Art der Repräsentation und Kodierung von Informationen darstellen. Diese Art der Darstellung wurde zunächst für den Navigationsraum entwickelt, doch wurden solche "Grid-Like" Kodierungsmuster auch für zweidimensionale konzeptuelle Räume gefunden. Ich untersuche eine „Persönlichkeitsfläche“, die auf zwei Eigenschaftenwörtern basiert, die die Achsen dieser Fläche bilden. Mit den verwendeten Analysemethoden finde ich keinen Hinweis auf eine "Grid-Like" Kodierung für diesen konzeptuellen Persönlichkeits-Raum.

Table of Contents

List of abbreviations

Psychology

NEO-FFI: NEO Five-Factor Inventory

RSA: Representational Similarity Analysis

IPIP: International Personality Item Pool

Neural

PFC: prefrontal cortex

fMRI: functional magnetic resonance imaging

GLM: General Linear Model

ROI: Region of Interest

mPFC: medial prefrontal cortex

vmPFC: ventromedial prefrontal cortex

dmPFC: dorsomedial prefrontal cortex

EC: Entorhinal Cortex

TPJ: temporo-parietal junction

STS: superior temporal sulcus

Computational Modeling

RL: Reinforcement Learning

RW: Rescorla-Wagner

PE: Prediction Error

RP: Reference Point

SSE: sum of squared errors

BIC: Bayesian Information Criterion

RDM: representational dissimilarity models

MVPA: multivariate pattern analysis

RSA: representational Similarity Analysis

List of publications

Frolichs, K., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications*, *13*(1), 1-18.

Rosenblau, G., Frolichs, K., & Korn, C. W. (2023). A neuro-computational social learning framework to facilitate transdiagnostic classification and treatment across psychiatric disorders. *Neuroscience & Biobehavioral Reviews*, 105181.

Frolichs, K., Rosenblau, G., Gläscher. J., & Korn, C. W. (**in prep**). Patterns of activation during personality representation.

Frolichs, K., & Korn, C. W. (**in prep**). Grid-like encoding during personality learning.

1. Introduction

1.1 General Overview

To make sense of an ever changing world, humans must be able to learn quickly and accurately. To do so, they need to take advantage of any structure there is in this often seemingly random world. Luckily, the world we inhabit is not completely random, statistical patterns (e.g., similarities between items within a category or similar behavior exhibited by a person across situations) allow for generalizations that can greatly aid learning. These generalizations exist across all domains of knowledge and learning and can very often be abstracted to similar shapes or structures (hereafter named knowledge structures).

It is these knowledge structures that allow for generalization during learning that I want to explore in this thesis. The domain in which this learning will take place is personality learning, even though the scope of this work expands beyond personality learning to (social) learning in general. The overarching question I try to answer in this thesis is if humans use knowledge structures when learning about other's personalities, and if so, how exactly they are applied both in behavior and within the brain. Over the course of three studies, I investigate multiple knowledge structures and their specific applications during personality learning.

The knowledge structures we are interested in all capture personality traits in a varying degree of detail and are based on decades of personality research. From this research there is the consensus that personality traits are a person's typical tendency to behave across a wide range of situations, that is relatively stable over time. Personality traits have been researched extensively and have been found to largely constitute a multidimensional structure where items within each dimension are more related than those between dimensions. It is these relations that I try to capture in the proposed knowledge structures.

The framework I largely follow in this thesis is that of David Marr, who proposed that in order to understand an information processing system it must be described at three levels of analysis. The first level, called "computational," is the problem the system is trying to solve i.e., the main question of this thesis. The other two levels then focus on how the system is solving this problem. Most of the work in this thesis focuses on the second level, algorithmic, which questions which processes or algorithms are at play to solve the problem posed at the first level. In the context of the third level, implementational, the question arises as to how the algorithm can be implemented

in a physical system, with the brain's neurons, their firing, and their network of connections playing a crucial role (more details on Marr's framework are provided in the next section).

Even though the questions of how personality learning is achieved are not novel, I, together with my coauthors, push the boundaries of recent methods and create novel paradigms based on already established methods. For example, in the first study, we use the very established reinforcement learning models. But instead of just using these models as they are, we add on the knowledge structure component which gives us new insights into learning as well as modeling. In the second study, we use a relatively new method called representational similarity analysis (RSA). This method uses representations of different types of data so that they get similar shapes and can thus be compared and analyzed. With this analysis we look for evidence for these knowledge structures in cortical activity patterns. The structures we model are some of the most complicated to date in the area of personality learning. Finally, in the third study, we use a very recent research topic and we go down to the third level of Marr's framework and take a look at a specific type of cell, namely, the grid-cell. We take this analysis one level of abstraction deeper than previous research and look for evidence for a conceptual personality trait space represented by these grid-cells.

1.2 Theoretical Background

1.2.1 Marr's Tri-Level hypothesis

In his 1982 book Vision, cognitive scientist David Marr (Marr, 2010) introduced his Tri-Level hypothesis that states that in order to understand an information-processing task three complementary levels of explanation have to be understood. These three levels are: 1) **computational**, what is the goal this computation should solve (e.g., in this thesis: how does the brain solve the personality learning problem), 2) **algorithmic**, what algorithms can implement these computations (e.g., an algorithm that uses prior personality information together with current experiences), and 3) **implementational**, how can hardware/ neurons carry out the computation (e.g., specific cells or a network of cells that perform these computations). This framework has been quite influential in the cognitive sciences (Lockwood et al., 2020; Love, 2015; Mitchell, 2006), and I will use it as well to structure my work in this thesis.

In this thesis, I take the information-processing system to be the brain. The first level (i.e., computational) encompasses the main hypothesis I try to answer in this thesis: How do humans learn about the personalities of others? The subsequent second and third levels are then explored from several perspectives. That is, I use multiple methods to investigate how the brain might solve the personality learning problem. The second level (i.e., algorithmic) is investigated in the first two studies and the third level (i.e., implementational) in the third and final study.

First, in study 1, I explore human personality learning using computational models based on reinforcement learning. Second, in study 2, I explore the neural underpinnings of the representations used in these models i.e., I will look at the activity patterns in the cortex during social learning. For the third level I shift my attention towards grid-cells (Bellmund et al., 2018). Grid- and place-cells have in recent years captured scientists' attention because of their specific coding patterns. In the third study, I will explore if personality is cortically represented on such a conceptual plane.

However, first I will have a look at personality psychology and the generalizations discovered in this research.

1.2.2 Personality Traits and its generalizations

Generalizing experience to guide novel decision making is a hallmark of intelligent behavior (Miller et al., 2002). In order to generalize one has to make use of statistical structures (e.g., generalizations). Because this thesis focuses on personality learning, the statistical structures that are used and investigated will all revolve around personality traits. Personality traits capture the differences between individuals in their typical tendency to behave. This behavior extends across a wide range of situations and over a relatively long period of time (Ashton, 2018; R. McCrae & Costa Jr., 2008). Knowledge about others' personality allows humans to make accurate predictions about their future patterns of behavior (Epstein, 1979). This predictive ability is automatic and is supported by the brain's representation of others' current mental states, which also includes their likely future states (Tamir & Thornton, 2018).

One of the main goals in personality research is to find accurate representations of personality traits that still capture its most important components. That is, to find some factors or dimensions that best represent the whole of personality in a distilled form. Most of the work in this regard has been the lexical approach (Ashton, 2018). In the lexical approach one takes all words that describe a person's personality (for a specific language) and, through questionnaires, find which of these words (or traits) correlate together. Throughout the years of research, there have been a multitude of ways of categorizing and ordering personality traits, what started with 12 factors (Cattell, 1943) was quickly whittled down to more manageable numbers. In the end, most studies found five personality factors. These Five-Factor Models (FFM) were introduced in the early 1960's (Tupes & Christal, 1992), and repeatedly found again by different researchers in the English (Goldberg, 1990; R. R. McCrae & Costa, 1987) and foreign languages (Peabody & De Raad, 2002). The five dimensions (also called the Big-5) that can summarize a person's personality are: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to experience. Even though later research with newfound computational power and in languages other than English revealed six factors (Ashton & Lee, 2007; K. Lee & Ashton, 2004; Saucier, 2009), the Big-5 has been widely accepted as the foundation of personality (Soto et al., 2015).

Based on these studies, I formulated several possibilities of personality generalizations that can potentially be used during personality learning. From these generalizations computational models are created that allow me to investigate which of these generalizations (or combinations of

generalizations) humans might use during personality learning. For this, I will use computational models, at which we will look next.

1.2.3 Computational Modeling

The need for computational models in psychology and neuroscience alongside more common statistical methods (e.g., t-tests and ANOVAs) has been proselytized in recent years (Farrell & Lewandowsky, 2018; Forstmann & Wagenmakers, 2015; Lockwood & Klein-Flügge, 2020; Wilson & Collins, 2019; L. Zhang et al., 2020). The main argument is that computational models offer conceptual clarity, i.e., when using math to communicate these often abstract ideas and theories, there is a lot less ambiguity than when only using words (Cushman, 2024). These models come in many different shapes and sizes and are really only limited to the researchers' imagination (and maybe the computational power they have available).

The goal of computational models in the domain of this thesis is to explain behavior. That is, I have an agent (animal or human) that exhibits a behavior (e.g., learning in a task) that I want to understand better. Computational models are then the mathematical equations that link the experimentally observable variables (e.g., task difficulty) to the agent's behavior (e.g., how quickly they learn or react) (Wilson & Collins, 2019). Most often multiple models are created, each of these models becomes an algorithmic hypothesis that explains the agent's behavior in a slightly different way. Finding which model explains the behavior best provides us with insights into the underlying processes of the behavior.

In this thesis, I focus on reinforcement learning (RL) models (Collins & Shenhav, 2022; D. Lee et al., 2012; Sutton & Barto, 1998). Based on classical conditioning (Pavlov, 1927), these models look at trial-by-trial variations in behavior to understand cognitive processes. This dependence on history (i.e., the experience on previous trials) makes them especially useful for modeling and understanding learning (Daw & Tobler, 2014). In detail, RL models use a prediction error (i.e., the difference between expectation and outcome) generated on each trial to update future estimates and thus reduce future prediction errors. This process can be described with a simple mathematical formula (see chapter 2.1) and encapsulates learning across a wide range of tasks (Dayan & Niv, 2008; Dunne & O'Doherty, 2013; Niv & Langdon, 2016; Ruff & Fehr, 2014).

Reinforcement learning models have been used to explain how humans learn from others (Diaconescu et al., 2017; Hill et al., 2016; Najar et al., 2020) and for others (Garvert et al., 2015).

Furthermore, studies have shown they can also model how humans learn about specific characteristics such as generosity, emotional states (Fareri et al., 2015; Jones et al., 2011; Zaki et al., 2016), single traits (Hartley & Somerville, 2015) or aspects of personality (i.e., trustworthiness) (Delgado et al., 2005; King-Casas et al., 2005). These studies only look at the current situation or a single dimension. However, as research into personality has shown (see section 1.2.2) human personality is everything but one-dimensional, the multidimensionality inherent in human personality (i.e., all five factors) should thus be captured in these computational models when one wants to accurately model learning (Jolly & Chang, 2019).

We therefore deemed it important to create models that exactly capture this multidimensionality in personalities and could thus use this factor information on different levels of detail during learning. We introduce a novel modeling paradigm that captures this. These models are based on the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972) and are expanded to work with several knowledge structures (i.e., structures that represent personality traits at different levels of detail). The models and our additions are given a simple introduction in chapter 2.1, and are discussed and applied in detail in chapter 3.

Explaining the behaviors during learning is only one side of the story. To get a full picture of human learning we also need to understand what processes are going on in the brain. Therefore, next, I want to have a look at the cortical processes during social learning and the regions I expect to be implicated.

1.2.4 Neural Representations of social learning

On a neural level the questions in this thesis are twofold. First, staying in the second level of Marr (i.e. algorithmic) I expect to find evidence for the brain resolving personality learning on the scale of representations. Specifically, the knowledge structures representing personality traits on different levels of detail (chapter 4). Second, descending to the third of Marr's levels (implementational) I want to explore grid-cells, specifically, in the light of this thesis whether there is grid-like encoding of personality trait space (chapter 5).

The regions I am most interested in are those implicated in the social brain (Frith, 2007). The areas of the social brain include the medial prefrontal cortex (mPFC) (Kanske, 2018; Saxe & Powell, 2006), medial parietal cortex, temporo-parietal junction (TPJ) and the superior temporal sulcus (STS) (Mitchell, 2006; Van Overwalle & Baetens, 2009). The social brain, particularly the

medial prefrontal cortex (mPFC), plays a crucial role in self-directed thinking and processing social information about others (Meyer et al., 2019; Mitchell et al., 2005; Van Overwalle, 2009). Furthermore, the mPFC is involved in inferring the mental states of others based on self-reflection (Meyer & Lieberman, 2018; Mitchell et al., 2005), and in integrating social information across time, allowing for reflection and representation of traits and norms (Van Overwalle, 2009). These findings highlight the importance of the mPFC in social cognition and its role in understanding and processing social information. For both studies the mPFC will therefore be the main region of interest.

1.2.4.1 Activity patterns of personality

The next study builds on the findings from the computational modeling. That is, the hypothesis that humans use knowledge structures of different granularities based on the Big-5 personality dimensions. Next, I want to expand on this research by exploring whether these knowledge structures are also represented in the cortex during a personality learning task.

As mentioned above, the main region of interest (ROI) for this experiment is the mPFC. The mPFC is instrumental in decoding mental states (Saxe et al., 2004) and in learning about these states through trial-by-trial feedback (Behrens et al., 2008; Gläscher et al., 2010; Joiner et al., 2017a; Korn et al., 2012; Korn, Fan, et al., 2014). Specifically, personality trait learning and social network structures have been implicated in the mPFC (Mitchell et al., 2004; Parkinson et al., 2017; Tamir et al., 2016; Thornton, Weaverdyck, Mildner, et al., 2019). Furthermore, there is strong evidence for different areas of the mPFC being responsible for self and other processing. Where the ventromedial prefrontal cortex (vmPFC) handles self-referential processing (Amodio & Frith, 2006; Sul et al., 2015; Wagner et al., 2012), the dorsomedial prefrontal cortex (dmPFC) handles other-referential processing (Saxe, 2006; Sul et al., 2015).

To investigate the existence of these knowledge structures in the cortical activity patterns, I use representational similarity analysis (RSA). RSA forms a subclass of multivariate pattern analysis (MVPA) (Kriegeskorte, Mur, & Bandettini, 2008) and is very commonly used in social neuroscience research (Freeman et al., 2018; Parkinson et al., 2014; Peer et al., 2021; Riberto et al., 2022; Stolier et al., 2020; Tamir et al., 2016; Thornton, Weaverdyck, Mildner, et al., 2019; Thornton, Weaverdyck, & Tamir, 2019; Thornton & Mitchell, 2017, 2018; Weaverdyck et al., 2021). Compared to standard fMRI analyses, which focus on mean levels of activity within a

region, MVPA examines patterns of activity across voxels within the population (Haxby et al., 2001). RSA, specifically, measures the degree to which these voxel patterns are similar across different experimental conditions (Kriegeskorte, Mur, & Bandettini, 2008). The biggest advantage of RSA is that it can compare different modalities (e.g., results from behavioral data with neural activity patterns (from different species)) (Kriegeskorte, Mur, Ruff, et al., 2008). It does this by not directly comparing the stimuli but rather a higher-order representation of each stimulus. Each stimulus' correspondence to the others is determined with a distance measure (Dimsdale-Zucker & Ranganath, 2018; Popal et al., 2019), which reveals how close each stimulus' representation is to the others.

There has been plenty of research using MVPA to look at representations (knowledge structures) of personality traits. Hassabis and colleagues gave participants four fictional characters who could be reliably classified based on two dimensions (agreeableness and extraversion) and found distinct regions that coded for these personality traits (Hassabis et al., 2014). Thornton and Mitchell (Thornton & Mitchell, 2017) used RSA to look at five different models of personality, ranging from two to five dimensions, and found evidence for their representations for all of them. However, based on our recent work (Frolichs et al., 2022) I expect humans to represent personality at even finer levels of detail than those found in this previous research. I therefore created several models of differing granularity (i.e., level of detail) to test whether the human cortex indeed represents these detailed representations of human personality when learning about others.

1.2.4.2 Grid-like encoding of personality space

In the first two studies, I look at behavioral and cortical generalizations based on the second level of Marr (algorithmic). That is, what algorithms or computations underlie the solving of the personality learning problem. In the final study I would like to take a look at a potential neural implementation (level 3 of Marr) of these generalizations in the form of grid-cells.

Grid cells and the accompanying place cells located in the entorhinal cortex (EC) and hippocampus have intrigued scientists for a number of years. Their firing patterns, revealed by single cell recordings, that tile arenas as rats move through them (Hafting et al., 2005; O'Keefe & Nadel, 1978) was first suggested by Tolman (Tolman, 1948) who hypothesized about their existence based on latent learning in the rats. That is, the grid- and place-cells represent the rat's current position in the arena and provide a context-invariant code to localize across similarly

structured environments (Behrens et al., 2018; Bush et al., 2015). This means that after creating a cognitive map of these arenas, the rats could find novel shortcuts within them. First evidence for grid-cells in the human entorhinal cortex came from an indirect analysis of fMRI data (Doeller et al., 2010) and later single cell recordings in the human entorhinal cortex confirmed their existence decisively (Jacobs et al., 2013).

Along with evidence for spatial representations, research has found that the mPFC, together with the hippocampus, generalizes knowledge for the task at hand (Mack et al., 2016). That is, the (v)mPFC updates existing knowledge representations during learning (Mack et al., 2020; Park et al., 2019) and the orbitofrontal cortex keeps track of currently relevant information in a cognitive map-like manner (Schuck et al., 2016). This means that potentially any type of information that can be projected on two dimensions can be represented by grid-cells in the mPFC (Behrens et al., 2018; Bellmund et al., 2018). Evidence for these representations has been bountiful with evidence ranging from pure abstract conceptual spaces (Constantinescu et al., 2016) to visual field (Nau et al., 2018) and olfactory (Bao et al., 2019) representations. Grid-like social representations were also suggested (Schafer & Schiller, 2018) and first evidence has been found (Liang et al., 2023; Park et al., 2021).

Because I hypothesized that the mPFC is the main region implicated in the generalizations from the personality literature in the previous chapter and evidence points towards the mPFC displaying grid-like coding for social representations, I hypothesize that a grid-like encoding of personality space exists in the mPFC. In detail, I expect this map to generalize along the dimensions from the Big-5. The conceptual personality trait space (hereafter called trait space) for this study consists of two dimensions along which people can be organized.

1.3 Thesis structure

My exploration of the above questions is executed in three separate studies. First, in **chapter 2**, I will pay specific attention to the three main methods used throughout my thesis. That is, computational modeling with a focus on the Rescorla-Wagner learning rule, representational similarity analysis, and grid-cell analysis. My goal is to give a non-exhaustive introduction to each method. To achieve this, I will give a high level overview of each of these methods accompanied by some code examples and figures. **Chapter 3**, will explore the strategies humans employ when learning about others' personalities. I introduce a novel modeling paradigm based on reinforcement learning to understand social learning across a wide range of tasks. **Chapter 4**, uses representational similarity analysis to examine neural firing patterns that we expect to reflect conceptual patterns about personality. I dive deeper into specific firing patterns in **Chapter 5** where we explore grid-like coding specifically for a personality trait space. Finally, the results and their implications are discussed in broad terms relating to current psychology, and (computational) neuroscience research in **Chapter 6**.

The perspective of this thesis is written in both singular and plural perspectives. Single for the sections specific for this thesis (i.e., chapters 1 (introduction), 2 (methods overview), and 6 (conclusion) and plural for the remaining chapters (i.e., chapters 3-5) which describe the studies. Because I can, in no good conscience, write these chapters from a singular perspective when the work in it comes from purely a collaborative effort.

2. Overview of the Methods – all studies

Significant portions of my time I have devoted to understanding and, in some cases, explaining specific methods that I have used in my research. This chapter is intended as a short primer on the three main methods that I have used and that will be used in the next three chapters of this thesis. First, I will focus on computational modeling of human behavior, in specific, the models that fall under the Rescorla-Wagner rule (Rescorla & Wagner, 1972) and the additions we have made to these models (Frolichs et al., 2022; Rosenblau et al., 2021). After that I will discuss representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), that aims to compare representations among different methodologies (e.g., behavioral, computational, and neural). Finally, I will look at a relatively new question in the cognitive sciences, namely the search for Grid-Cells (Hafting et al., 2005), specifically for tasks that are not in physical space (i.e., conceptual spaces) (Constantinescu et al., 2016), using functional magnetic resonance imaging (fMRI) (Doeller et al., 2010).

My intention for this chapter is to give someone a head-start that I sometimes wish I had on some of these topics.

Code for all the examples and figures has been written in Matlab and has been uploaded to Github (https://github.com/kFrolichs/Thesis). In some cases these figures have animations and thus add a bit more information when run from Matlab compared to the non-animated (digital) paper versions presented in this thesis.

2.1 Computational Modeling

The success of mathematical models to explain phenomena in our environment cannot be overstated. Take the Copernican model of our solar system with the sun in the center and the planets revolving around it. The exact movements of which would later be described by applying Newton's laws. By now mathematical models are ubiquitous to describe and predict anything from today's weather to tomorrow's stock market. When using computers to calculate and fit these often complex mathematical models it becomes computational modeling. In the cognitive sciences we use these computational models to describe and predict people's behavior (Farrell & Lewandowsky, 2018). In short, computational models add another tool in our toolbox that grants a new perspective separate from traditional statistics, especially because we can look at how behavior functions in time, that is, figure out the algorithms that underlie behavior (Wilson & Collins, 2019). There are numerous modeling frameworks out there, in our work we have focused on reinforcement learning models, and thus only these will be described briefly below.

2.1.1 Reinforcement Learning Models

Introduced by Sutton & Barto (Sutton & Barto, 1998), Reinforcement Learning (RL) models, that learn by trial and error feedback from their environment, have quickly grown to be in use for many functions. Most famously used by artificial agents in Deepmind's AlphaGo (Silver et al., 2016) and Protein folder (Senior et al., 2020). Maybe slightly less well-known, these models have been used widely in the cognitive sciences in an attempt to explain learning and decision making (Dayan & Niv, 2008; Dunne & O'Doherty, 2013; L. Zhang et al., 2020). Most strikingly, these algorithms correspond well to the firing patterns of dopaminergic neurons in the midbrain (Daw & Doya, 2006; P. Montague et al., 1996; Schultz et al., 1997). Since these first seminal studies these algorithms have found wide use in the cognitive sciences. In this chapter I will only focus on a specific learning rule, namely, the Rescorla Wagner rule (Rescorla & Wagner, 1972). Furthermore, I will only look at model comparison i.e., figure out which of a set models best describes the behavioral data.

2.1.1.1 The Rescorla-Wagner learning rule

The Rescorla-Wagner learning rule (Rescorla & Wagner, 1972) was invented to explain blocking, a specific phenomenon observed in classical conditioning. Blocking happens when multiple stimuli are presented. The first stimulus, who's relationship with the reward was learned first, blocks the learning about any of the new stimuli's relationships with the reward.

Let's first have a look at how the RW-rule can explain classical conditioning (i.e., Pavlov's dogs) (Pavlov, 1927). In the most well-known experiment, Pavlov measured dogs' salivating response to hearing a bell. Interestingly, this bell was paired with the dogs receiving food shortly afterwards. Normally, dogs salivate when receiving food but after repeated exposure the dogs already began salivating at the sound of the bell (they had become conditioned).

The RW-rule captures this process in a simple algorithm that updates its estimate about the environment based on feedback from this environment. In the case of the dogs there is an estimate (hearing a belling sound means nothing) and an outcome (after the belling sound I received food). In numbers we could express this in the following way. The expectation of food can be expressed from 0 to 1 (0 meaning not expecting food at all and 1 indicates that one expects food 100% of the time). Furthermore, we need a value (between 0 and 1) that represents how fast the dog will learn, conveniently named the learning rate. A standard value for this is 0.05, so we will stick with this as well as with its commonly used symbol, the Greek letter α. What the RW-rule will explain is how the expectation of food changes over time after repeated pairing of the belling sound together with receiving food. One small detail to add is that we will treat time as if it was discrete i.e., as if it was made of steps in time, where each step in time is indicated with t, where t means the first time step, $t + 1$ the second, $t + 2$ the third, and so on. Every trial will be a new time step.

First we need to calculate the prediction error (PE, equation 1), that is the difference between the expectation and the outcome. Initially the dog expects nothing but then receives food. This results in a prediction error of 1: $1 - 0 = 1$. Next, the expectation gets updated for the next trial, this takes the current expectation and adds the PE multiplied by the learning rate (equation 2). Resulting in a new expectation of $0.05 = 0 + 1 \cdot 0.05$. This means that for the next trial, when hearing the bell, the dog has a small expectation of food (**Figure 2.1A**).

$$PE = Outcome - Expectation \quad (1)$$

$$Expectation_{t+1} = Expectation_t + PE \times \alpha \qquad (2)$$

On repeated exposure to the bell and food, the expectation of food will slowly increase (**Figure 2.1B**). When the food does not show up the expectation will decrease a little bit (**Figure 2.1C**).

Since this is a simple equation with only one free parameter (the learning rate), we can only make it change its "behavior" a little bit. That is, we can only change the value of the learning rate (**Figure 2.1D**). Doing so changes how quickly the dog will learn. This can also help us distinguish dogs based on their learning rate.

The example of the dog and the bell can, of course, be exchanged for any repeated learning paradigm such as: playing on slot machines (e.g., bandit tasks) (Daw et al., 2006; Lau & Glimcher, 2005; Sugrue et al., 2004), social interactions (FeldmanHall et al., 2018) and preference learning (Rosenblau et al., 2018, 2021).

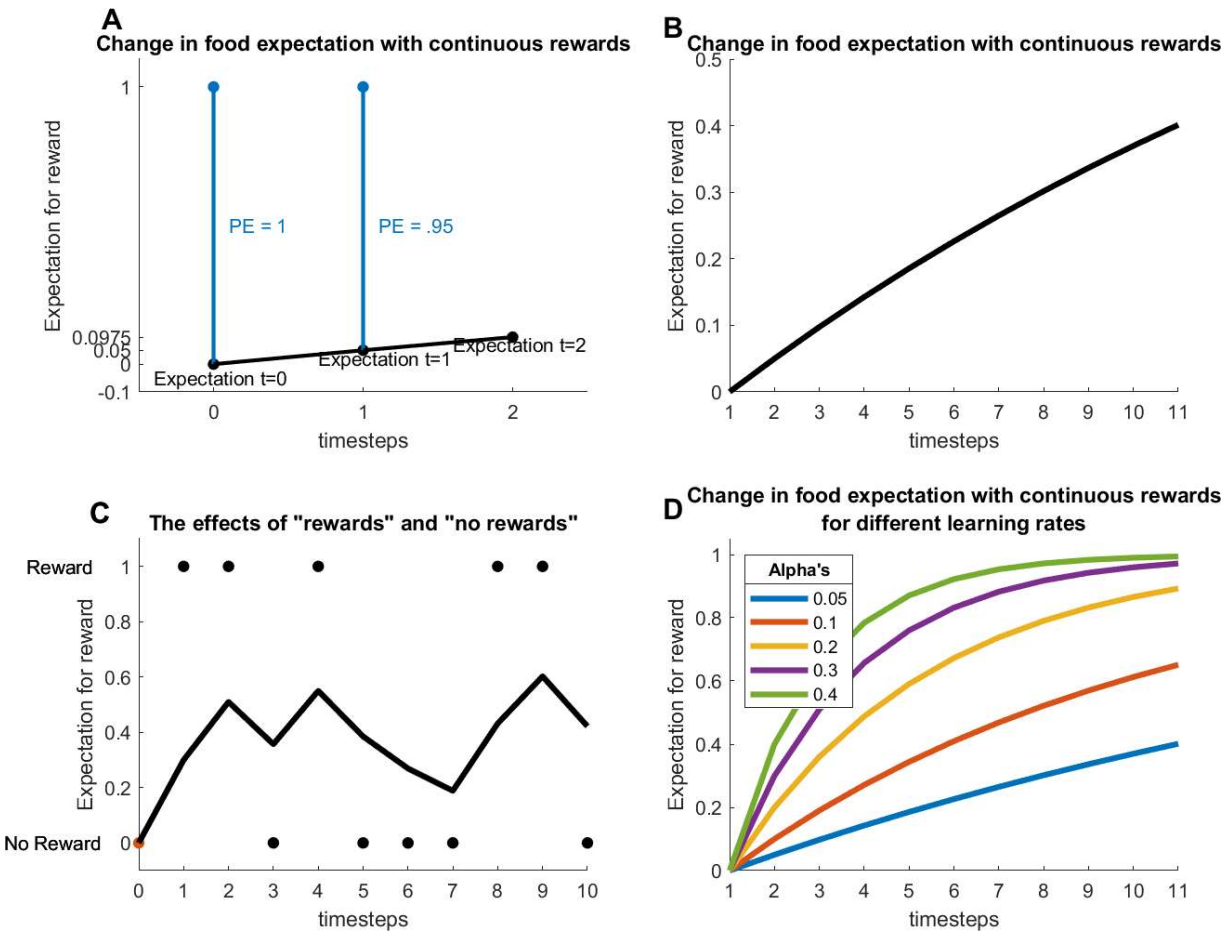


Figure 2.1. Exploration of the Rescorla-Wagner learning rule.

a) Starting with no reward expectation (i.e., 0) and a learning rate (α) of 0.05 this agent slowly updates to start expecting rewards. **b**) Shows the increase in reward expectations over a longer time. **c**) When an agent is not continuously rewarded their expectations can also lower again. **d**) The learning rate determines how quickly an agent updates. The higher the learning rate the quicker the updating.

2.1.1.2 Increasing the complexity of the models

Having the standard RW-rule is great, as shown before it can explain varying phenomena. However, we would like to dig a little deeper and test more complex behavior. In this case, how humans learn about others' personalities. Specifically, what knowledge structures they use when learning about others. That is, we present participants with a task where they have to learn about strangers' personalities, by repeatedly asking them to estimate how they would score on a certain trait word (e.g., diligent). To do this we can keep the RW-rule as our foundation but add variables and data to these standard equations. This way, every model will become a specific hypothesis of strategies that humans can employ when learning about others.

Adding these complexities is relatively straightforward. A lot of research points towards a 5 factor model of personality traits (Goldberg, 1990; R. McCrae & Costa Jr., 2008). That is, personality traits can be grouped into 5 separate factors that contain trait words that are connected to each other. For our first model we hypothesize that people only learn by updating a single value for each factor. That is, they have 5 values in memory, one for each factor, and only update these values during learning. So for each trait word they get presented, they first categorize it into the correct factor and then get that factor's value, when receiving feedback they also only update this factor value (equation 3). Where F indicates the factor the current trait word belongs to.

$$Expectation_{(t+1,F)} = Expectation_{(t,F)} + \alpha \times PE \qquad (3)$$

However, usually when we learn about someone we do not start from a blank slate. We have already encountered numerous people before and therefore have opinions or ideas about others available. In this case we expect people to already have formed ideas about the population (students) they are learning about. We took a separate sample of people and asked them about the personality traits of the average student. We can now expand the model from equation 3 by adding this new source of information hereafter named the reference point (RP). Furthermore, we want to find out how

humans integrate the information gained from their environment with their preconceived knowledge. We thus add a weighting between the standard factor learning and the RP in our new equation (equation 4). This weighting is indicated by the Greek symbol: γ. Thus, if the γ is 1, one will only use the RP, if it is 0 one will just be using equation 3.

$$Expectation_{(t+1,F)} = \gamma \times RP + (1-\gamma) \times (Expectation_{(t,F)} + \alpha \times PE) \qquad (4)$$

2.1.1.3 Applying the model to real data

Now that we have some models that represent possible strategies that humans can use. We want to apply it to real-life data. This works using model fitting. Model fitting means we change the free parameter(s) to get the best fit of the model to the data. First, we need to know what fit is. In this case, we want to fit our model to our participant data. Let's assume participant data are ratings and our model also outputs ratings, we can visualize this as two separate lines (**Figure 2.2A, B**), the more closely they resemble each other the better their fit is i.e., **Figure 2.2A** shows a bad fit between the lines and **Figure 2.2B** a good fit.

We would like more than just visually comparing lines so we use a metric called the sum of squared errors (SSE, equation 5). The sum of squared errors looks at the difference (errors) between the two lines at every point on the axis, squares them, and then adds them up (**Figure 2.2C, D**). The squaring is done for two reasons, first it changes all values to positives, second, it punishes more for larger errors. With the SSE we can now capture the fit of the model to participants' behavior in one number. The SSE is just one of many metrics that can be used, each with their own pros and cons.

$$SSE = \sum_{t=1}^{n} (data_t - estimates_t)^2 \qquad (5)$$

Now we can determine what value of the free parameter(s) result(s) in the best fit by stepwise changing the value of the free parameter and seeing its effects on the fit (Figure 2E). In these simple models with simulated data that has no noise, this can be done by hand but when working with real life data we usually use an algorithm to do it for us. In our case we use the Nelder and Mead simplex algorithm that is already implemented in Matlab (*fminsearch*). When we run our model on this we get the following result (**Figure 2.2F**).

Furthermore, we can compare different models' fit with the SSE to see which of these models best explain the analyzed data.

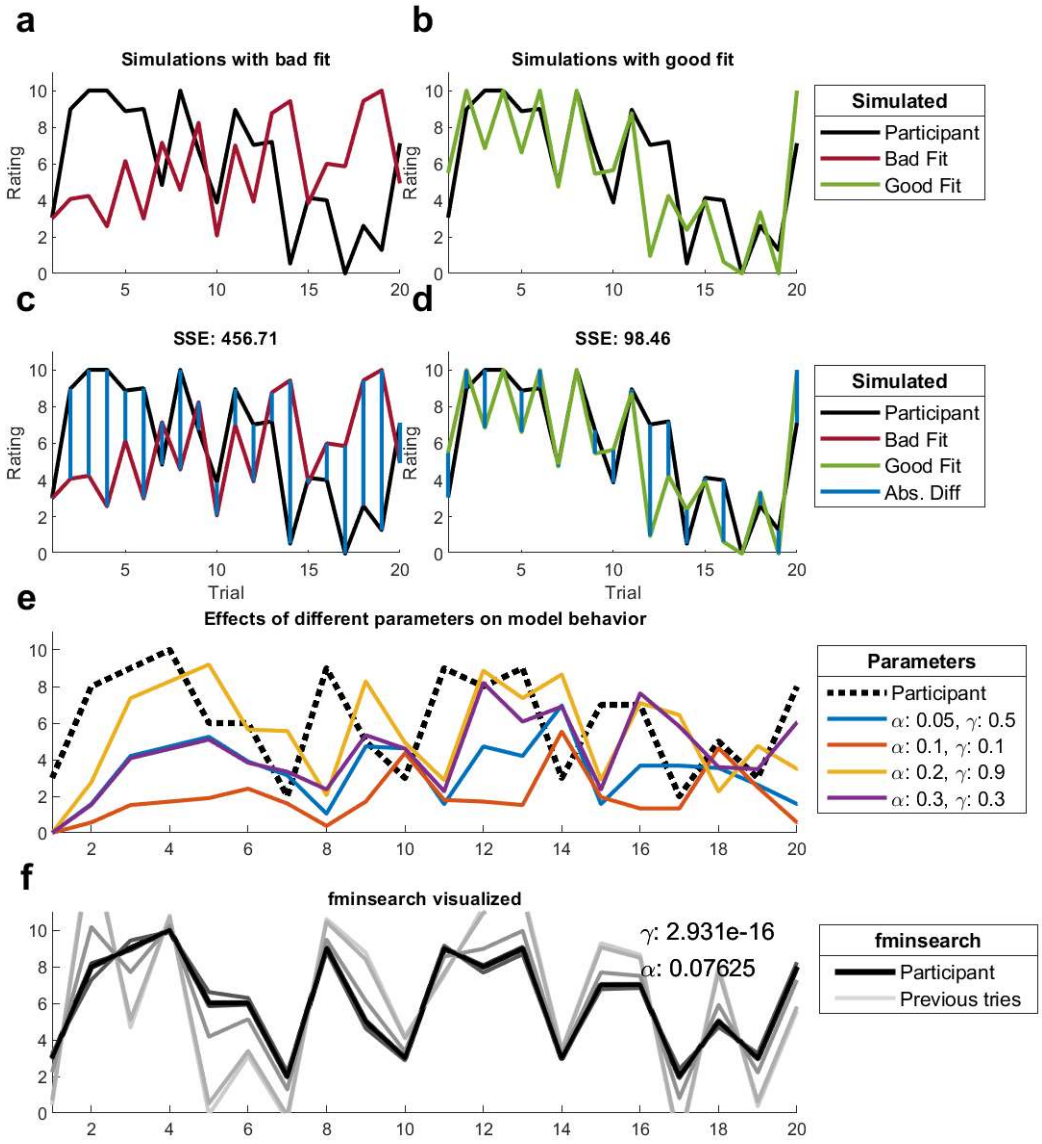


Figure 2.2. Overview of model behavior and model fitting

The black line always shows choices from a simulated participant. **a**) Shows a bad fitting model (red line). **b**) Shows a better fitting model (green line). **c**) and **d**) Show how well the models fit to the data using the sum of squared errors (SSE). **e**) in a model with two free parameters (α and γ) the model can show a lot of different "behaviors" i.e., the different colored lines all originate from the same model but with different free parameters. **f**) The fitting algorithm will systematically look through different free parameter values to iteratively find the parameters that produce the best fit.

We can now create and determine the fit of any model we can think about, but one last step still remains. It is unfair to just compare the models' fit like this because the models are not equal in their complexity. When we compare the models from equation 3 and 4 we see that model 4 has more free parameters. We need to punish for this, luckily there are also algorithms for this, in our case we will use the Bayesian Information Criterion (BIC) (equation 6). That uses the SSE we calculated before but also the number of free parameters (k) and the number of trials (n).

$$BIC = n \times ln(\frac{SSE}{n}) + k \times ln(n) \qquad (6)$$

This BIC value gives us a fair estimate of the fit of the model controlled for its complexity. This means we can compare models fairly now using their BIC values.

2.1.1.4 Model Validation

With the modeling hopefully a little bit clearer we can now focus on some extra steps that are wise to take during the whole procedure. These steps are discussed in more detail in the following papers (Palminteri et al., 2017; Wilson & Collins, 2019; L. Zhang et al., 2020). Here we will focus in particular on the confusion matrix and parameter recovery. In both cases we want to make sure our models behave the way we expect them to behave. To do so we will simulate data, this way we know exactly what data the models receive and thus how it is processed.

We calculate the confusion matrix to make sure our models are distinguishable, that means, we determine if their behavior is different enough so that we can differentiate them from another. If we cannot do this our whole modeling procedure is invalidated since we could be drawing conclusions based on one model even though another (that stands for a completely different hypothesis) would give us the same results. The end result of the confusion matrix looks like **Figure 2.3D**, this symmetrical matrix shows how often data that was simulated by one model was also determined as the best fit by our modeling procedure (0 being never, 1 always). Ideally this will result in an identity matrix (i.e., a diagonal of 1's, which indicates that the model that simulated the data was also the one that fitted it best). However, in reality these numbers are never perfect so we are usually happy with numbers in the high 0.90's.

We simulate the data by doing the opposite from what we normally do. That is, we give the models data from a task we created ourselves together with some values for the free parameters.

To get the parameter values we can simply use a random number generator to either give us numbers between the bounds of the parameter or between the observed bounds if we have already fitted models on participants' data. The model output is then regarded as our data (i.e., you can see this as if we have a participant (the model) for which we know the exact strategy and free parameter values). We do the same procedure for every model that we have and subsequently fit all the models on this generated data to create the confusion matrix.

For parameter recovery we want to know how well the models can retrieve the correct parameters from data. This largely follows the same procedure as the confusion matrix where we know what we gave as input and can therefore check it with the output, in an ideal situation recovering the same parameters as those that we used to generate the data. Similar to the confusion matrix we simulate data by pretending the models are participants. Now we keep track of the parameters that were used to simulate the data and we only fit the models on their own generated data. We visualize the parameter recovery with a scatter plot, where ideally inputted and generated parameters fall on a straight line (**Figure 2.3A-C**).

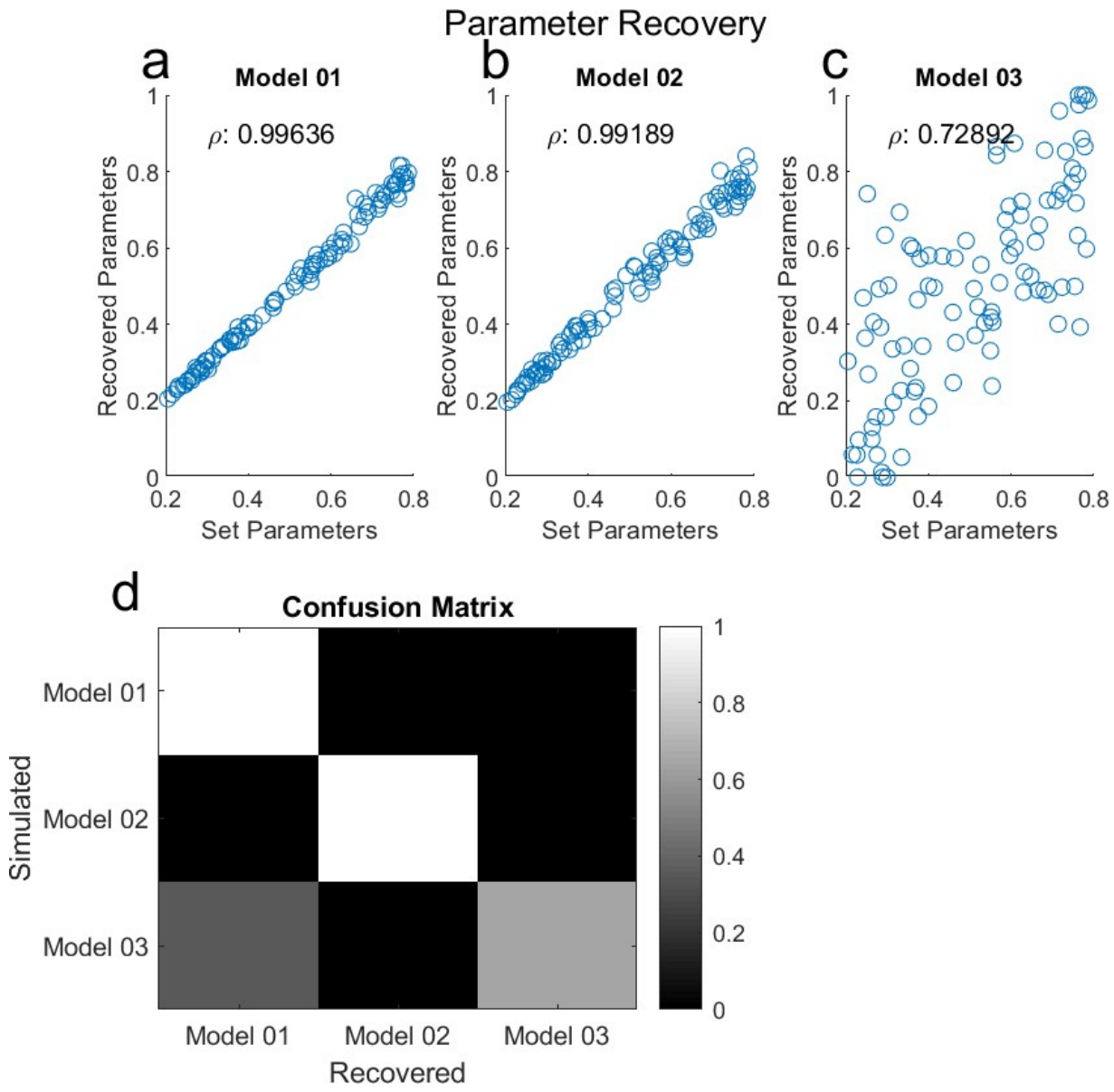


Figure 2.3. Parameter recovery and confusion matrix on three different models.

a-c) Shows the parameter recovery results from three different models. Models 1 and 2 are exemplary and model 3 would be acceptable. **d**) Shows results from the confusion matrix. Again, models 1 and 2 are exemplary. The results from model 3 show that this model frequently gets confused with model 1.

Resources

Matlab code for all examples has been uploaded to Github:
https://github.com/kFrolichs/Thesis/tree/main/2.1-Computational_Models

A Python notebook hosted in Google Colab from a 3-day course on computational modeling we taught in the summer of 2022 can be found here: https://tinyurl.com/5n6ep8md

2.2 Representational Similarity Analysis

In this section, I would like to give a short overview of the base ideas of representation similarity analysis (RSA). Similar to the previous section the text will be accompanied by MATLAB code including toy problems to further illustrate the specific steps in this analysis pipeline. Representational Similarity Analysis is a subcategory of multivariate pattern analysis (MVPA). MVPA in fMRI makes use of the fact that activity patterns in voxels carry information that can be en- and decoded (Haxby et al., 2001). By converting these activity patterns to high-dimensional representational vector spaces we can use machine learning methods to define decision boundaries. These decision boundaries allow us to answer questions such as what information is represented in a specific region i.e., distinct patterns of activity that allow distinctions among brain states (Haxby et al., 2014).

RSA analyzes the similarity between response patterns as distances in the aforementioned representational space. That is, it is a statistical technique based on analyzing second-order isomorphisms (i.e., similarities). That means instead of directly measuring the relationship between two measures it first computes similarity measures within measures and then compares these similarities to each other. These distances can be calculated using varying methods such as Euclidean, correlation, or Cosine (Walther et al., 2016). Because it relies on these computed similarities and not direct measures it is a very flexible analysis that can compare vastly different measures (e.g., brain activity and behavior, even between different species) (Kriegeskorte, Mur, & Bandettini, 2008).

In our case we want to find out what knowledge structures are employed by the cortex when learning about others personality traits. That is, we want to compare similarities between patterns of neural activity with multiple conceptual models - representing different knowledge structures - to find which of these structures are represented in the cortical activity patterns.

2.2.1 RSA step-by-step

In the following sections we will create and analyze a toy dataset. Illustrating some parts of the analysis we have performed for study 2. In brief, we want to find-out what personality representations are represented in the cortex during personality learning. Specifically, we look at two different granularities of the personality representations. First, a coarse representation that only

reflects the Big–5 factors that each trait can belong to. Second, a fine representation that represents each trait word separately together with its similarities to all the other trait words. Furthermore, we will only use correlations as the distance metric but this can easily be substituted by any of the other common distance metrics mentioned above.

2.2.1.1 Creating Toy Data - Model & Data RDM

First, we want to create some artificial data to work with. For the "task", we will present 50 trait words to each participant. These 50 trait words can be subdivided into 5 (Big-5) factors (Goldberg, 1990; R. R. McCrae & Costa, 1987) and we want the 10 items within the factors to correlate higher than between the factors (**Figure 2.4A**).

These similarities should be reflected in the model representational dissimilarity matrices (RDMs) that we are creating. These are square matrices that compare the dissimilarities between items. The dissimilarity is used because this makes more intuitive sense from a distance perspective (i.e., a dissimilarity of 0 means "no distance" and thus "the same", whereas a similarity of 1, which carries the same meaning, does not give this intuitive advantage). The coarse model (**Figure 2.4B**), which consists of zeros and ones, only takes into account what factor the trait word is part of. In the perspective of our RSA analysis this means that we expect the activity patterns within factors to be very similar (i.e., 0's: no distance) and between factors very dissimilar (i.e., 1's: maximum distance). The fine model in this case is the previously created correlation matrix subtracted from 1 to create a dissimilarity matrix (**Figure 2.4C**).

These model RDMs are our hypotheses i.e., these are the expectations we have for the patterns of activity in the cortex. We will compare these model RDMs with the data RDMs to see which of these models can describe the data well.

The final step in creating artificial data is to create the neural activity patterns. We will keep it very simple and will not add any noise specific to fMRI outside of some normally distributed random noise. The activity patterns of some of these neurons are displayed and already show some overlap between neurons for some factors (**Figure 2.4D**).

2.2.1.3 Calculate data RDM's

To calculate the data RDMs we first need to calculate beta maps using a standard general linear model (GLM). This would normally also correct for the many variations of noise present in fMRI

data (e.g., movement). Our data is kept pretty clean though so we will only include an intercept and a single regressor for each stimulus separately for our GLM. This GLM will return a beta map for each stimulus. That is, the corrected activity for each voxel for each stimulus. From these beta maps we can calculate the representational space of how this region of cortex responds to each image (**Figure 2.4E**). That is, each value within the data RDM represents the correlation between activity patterns for two stimuli. For example, we would expect that the activity patterns of two trait words that are more related such as *generous* and *kind* are more correlated than the activity patterns of two unrelated trait words such as *generous* and *diligent*.

Usually this is not done for the whole brain but rather within a region of interest (ROI), you specify beforehand what regions of the brain you are interested in and only look at the voxels within this specific area. So look at a specific ROI, take only the voxels from within this ROI and calculate the RDM from here.

2.2.1.4 Calculate model and data similarities

Now that we have both the data (**Figure 2.4E**) and model RDMs (**Figure 2.4B,C**) (and they have the same shape i.e., square matrices the size of the number of stimuli) we can calculate if the brain might process information like we expected based on the model RDMs. To compare these RDMs (the data RDM with each of the model RDMs) we calculate the Spearman's correlation coefficient. We use Spearman over Pearson's coefficient because the stimuli might not have linear relationships. Other methods for determining the similarity between RDMs are Kendall's Tau alpha (Walther et al., 2016) or permutation testing (Dimsdale-Zucker & Ranganath, 2018; Etzel & Braver, 2013). But since the Spearman correlation is native in Matlab we will use that for simplicity's sake. After receiving the correlation coefficients we standardize the coefficients using a Fisher's F-to-z transformation. These standardized values will then be subjected to a one-sample t-test (over participants) (**Figure 2.4F-H**).

2.2.1.5 Searchlight analysis

Another option is to use a searchlight, this is when one creates a specific shape (usually a sphere) and moves it through the brain. At each position you calculate the RDM from that specific set of voxels and compare it with one (or more) model RDM(s) using any of the above-mentioned distance metrics, this value gets assigned to the center voxel of the sphere. This way, you create a

whole brain map where each voxel is a correlation coefficient with regard to a specific model RDM. This brain map then reflects where in the cortex (if anywhere) patterns of activity correlate with the tested model RDM.

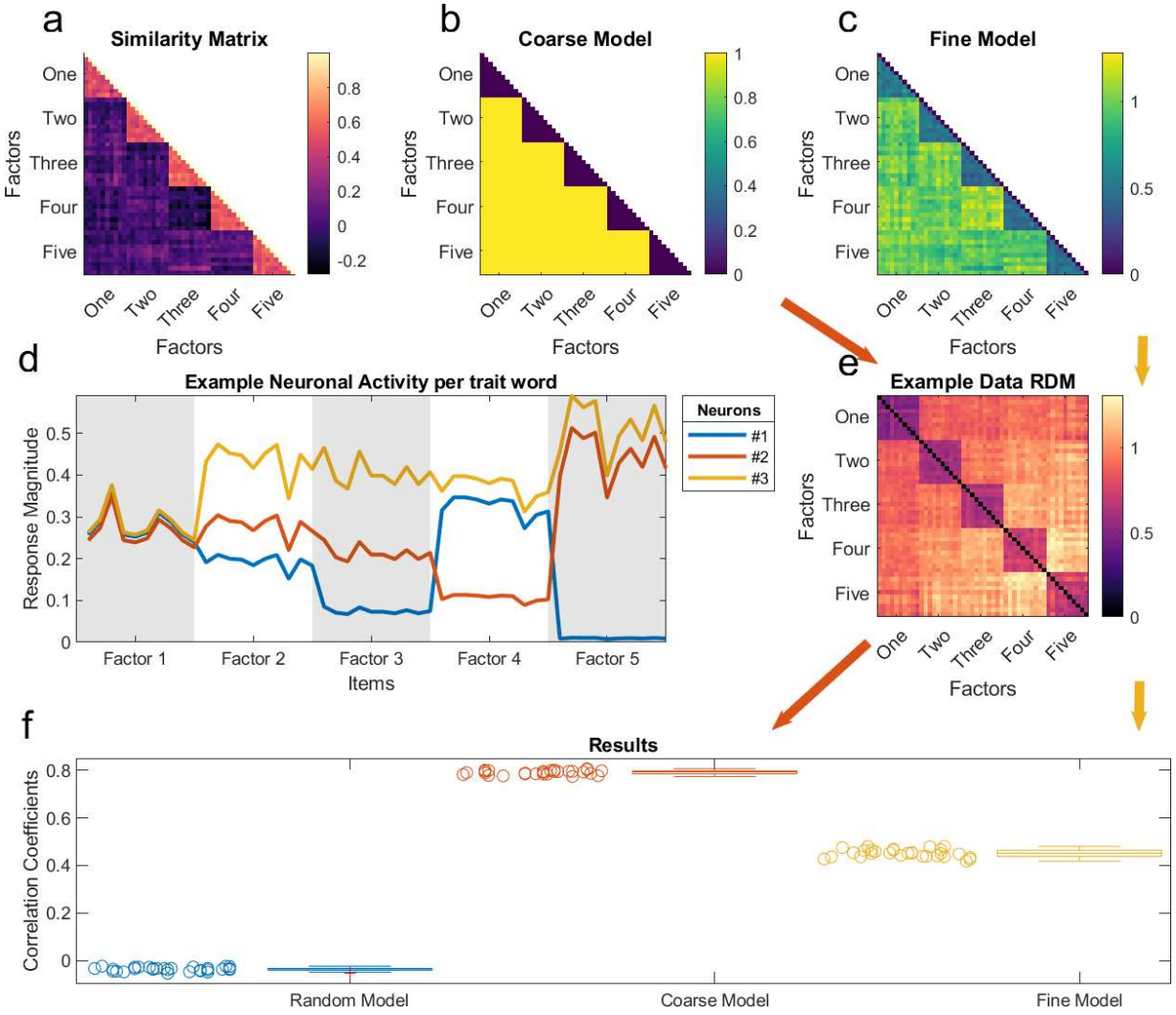


Figure 2.4. Overview of some of the main components of an RSA analysis.

All data represented in this figure is simulated for illustrative purposes.

a) The similarity matrix that was simulated for this figure. Items within a factor are more correlated than between. **b**) The coarse RDM assumes that activity patterns from items within a factor are exactly the same and the patterns between factors completely dissimilar. **c**) The fine RDM is like the similarity matrix (a) but then subtracted from 1 to create a dissimilarity matrix. **d**) Some selected neurons' activity patterns. These neurons show similar activity for the items 10-20. The gray patch highlights the different factors. **e**) The data RDM we calculated from our data. There is a very clear pattern here that normally would not be so obvious. **f**) Results from the comparison between the data RDM and three model RDMs. We normally would use a one-sample t-test compared to 0 to test for significance. First, a model RDM

with random numbers, shows low to no correlations as expected. **T**he coarse model explains the data very well. This is of course expected from our simulated data with very little noise. Similarly to the coarse model, the fine model also explains the data well. These results would lead us to conclude that both the coarse and fine model explain patterns of activity significantly well in this data.

Resources

Code examples, and all the code for the figures can be found on Github: https://github.com/kFrolichs/Thesis/tree/main/2.2-Representational_Similarity_Analysis

2.3 Grid-Cell Analysis

In the 50s, Tolman (Tolman, 1948) suggested that rats might possess something like a cognitive map of their environment. He hypothesized this because rats showed latent learning (i.e., learned about an arena even when not directly interacting with everything). Later, work by O'Keefe and Nadel (O'Keefe & Nadel, 1978) suggested that this could take place in the hippocampus by so-called place cells. These place cells were named for their property of firing only in a specific location in the arena. Later work using single cell recordings from the rat entorhinal cortex as they were exploring an arena (Hafting et al., 2005). Found grid cells that showed activity in a grid like fashion, tiling the floor in hexagonal patterns. Concrete evidence for human grid-cells came from invasive recordings during virtual arena walking (Jacobs et al., 2013), this was after a method was developed that allowed the search for grid-like activity using fMRI also when engaged in virtual arena walking (Doeller et al., 2010). Since humans not only navigate actual space but also conceptual spaces, this was explored next (Constantinescu et al., 2016). Since then there has been a steady supply of evidence for grid-like activity in several domains (Bellmund et al., 2016; Nau et al., 2018; Park et al., 2021). In this primer I will focus on explaining the original methods used to find grid-like activity using fMRI.

We can find grid-cells with fMRI because of two main reasons. First, the orientation of the grid-cells seems to be constant across cells. The second reason has to do with the so-called conjunctive cells. These show modulation of their activity based on the direction in which one is moving. Because of this there are noticeable changes in activity patterns based on whether one is aligned (i.e., hitting the cells consistently) or misaligned with the grid (not hitting the cells consistently) (**Figure 2.5A,B**).

However, the origin (grid angle) is different for everyone, so we first have to determine that before we can look for the grid-like activity in the cortex. In this short primer I will only look at the main fMRI analysis that we used in our project.

2.3.2 Grid-Cells in fMRI data step-by-step

The scanner task should have some form of movement (i.e., real or conceptual). If there is grid-like coding, movement through this space will evoke it. The activity will depend on the angle the

movement is in relative to the arena. That is, we expect differences in response magnitude based on the angle of the movement.

The normal analysis procedure is to split the data in two. With the first half of the data we find the individual grid-angle of each participant (i.e., the baseline 0° angle that is different for each person). We then use the second half of the data to confirm the accuracy of this grid-angle. This procedure is done with two separate GLMs explained in further detail below. But first we will generate some data that has activity patterns that reflect grid-like coding.

2.3.2.1 Create toy data

We want to create artificial neural data that reflects grid-like coding. That is, a signal that has 6 peaks spaced at equal intervals (i.e., every 60°) (**Figure 2.5A, B**). In our case we will give it a peak angle of -8° so we can see if our next analysis steps correctly return this peak angle (**Figure 2.6A**). **Figure 2.6B** shows the same pattern as in A. This was added to show that it contains the same (type of) information as the design matrix. Normally the exact angle for these peaks differs for each person, it is thus important to know we can accurately estimate it. We attempt to find this peak angle with the first GLM.

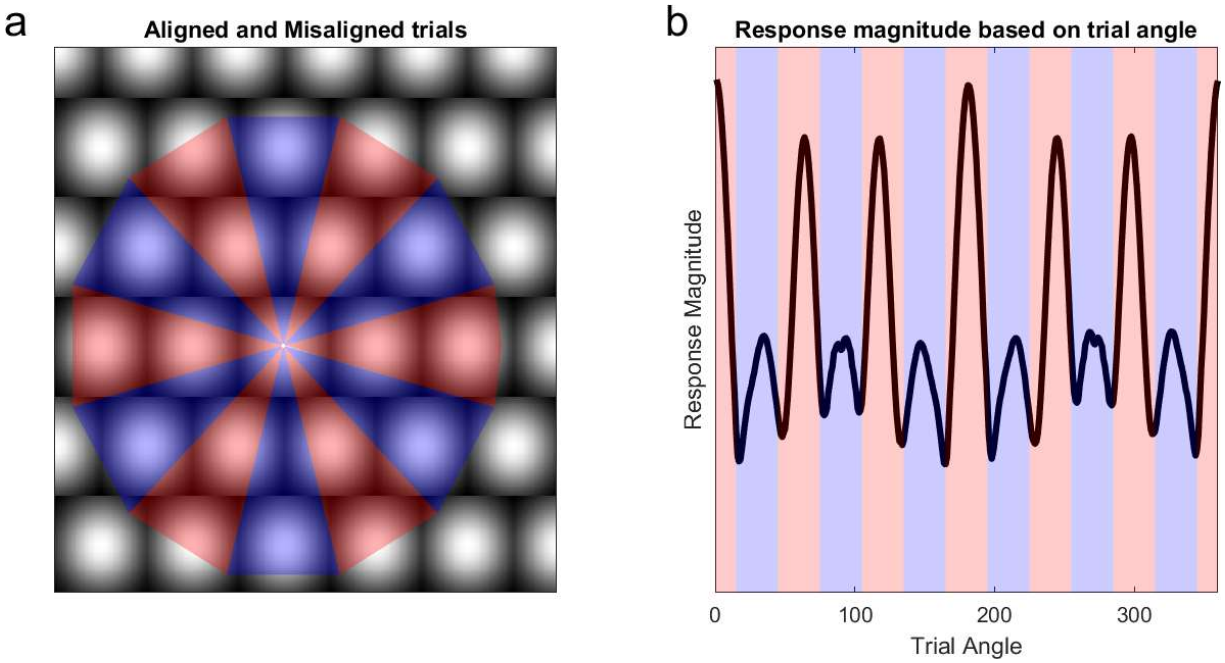


Figure 2.5. Grid-like coding.

a) the grid-cell as white dots tiling the arena. The overlaid colors indicate whether one would be aligned with the grid (red) or misaligned (blue). **b**) the black line indicates what the potential activity could look like for running aligned vs misaligned.

2.3.2.2 First GLM - determining the grid-angle

We attempt to find the individual grid angle by creating two parametrically modulated regressors. These parametric modulators contain the sine or cosine for the angle of each specific trial (θ_t) respectively [$sin(6 \cdot \theta_t)$ and $cos(6 \cdot \theta_t)$]. We multiply times six because we expect the signal to be sixfold. That is, have six peaks in the data. Since sine and cosine are orthogonal we can create any signal by combining these two. We create a design matrix with the sine and cosine of the trial angles together with an intercept (**Figure 2.6C**) and calculate the resulting betas with a standard GLM. From these beta's we can calculate the personal grid angle φ with the following formula: $arctan(\beta_{sin} + \beta_{cos})/6$. In this case it returns a φ of -8.3246°, which is very similar to the angle used to create the data (**Figure 2.6D**). This indicates that our first analysis functioned as intended.

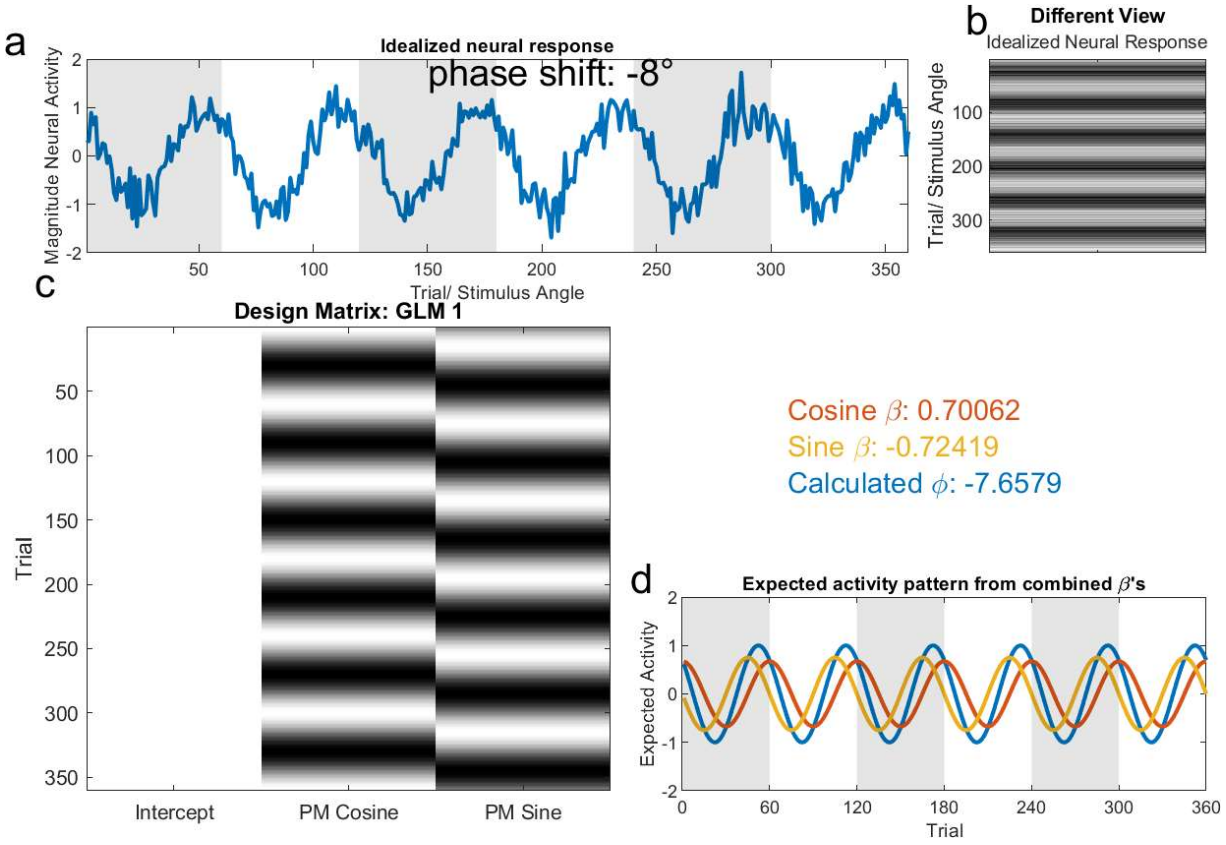


Figure 2.6. Grid-like coding and first GLM to find the personal grid-angle.

a) The neuronal activity pattern that we simulated has a phase shift of -8°. That is, it has peaks at 52°, 112° and the following intervals of 60°. **b**) Shows the pattern in (c) in a different view that's more alike that of the design matrix. **c**) The design matrix consists of an intercept (white) and two parametrically modulated regressors one for each the sine and cosine of the trial angles. **d**) From the beta's we receive for the two parametrically modulated regressors we can calculate this participants' individual grid-angle. This grid-angle is calculated to be -8.3246° which is pretty close to our original -8°, indicating that our methods worked successfully.

2.3.2.3 Second GLM - Consistency check

Now that we have the personal grid-angle we can check if we find this same angle on the second half of the data. There are a number of methods of doing this and I will highlight three related analyses.

In method one we create two regressors, one regressor contains the aligned and the other the misaligned trials (**Figure 2.5A,B**). Whether a trial is aligned or misaligned is determined by taking the participants' unique grid-angle calculated in the first GLM and adding ± 15° (i.e., a bin of 30° for both aligned and misaligned trials, together making up 1/6th (60°) of the available movement angles). These angles along with the 60° multiples will be the aligned trials (**Figure 2.5A**, red patches), the other trials will be the misaligned trials (**Figure 2.5A**, blue patches). In our example the trials are ordered from 1°-360° which creates a checkered pattern in the design matrix (**Figure 2.7A**). **Figure 2.7B** shows how these two regressors function in an ideal situation on the neural data, resulting in very clear evidence for grid-like activity in **Figure 2.7C**.

In the second method we will take the same steps as in the first method. That is, determine which trials are aligned and which misaligned based on the participants' unique grid-angle. However, instead of lumping all aligned and misaligned trials together we keep them separated. This results in 12 regressors (6 aligned and 6 misaligned) (**Figure 2.7D**,**E**). **Figure 2.7C**, shows how results would look from near perfect data, in reality it will be a lot more noisy.

In the third method we will create one parametrically modulated regressor with values based on the grid-angle (φ) from the first GLM along with every trial angle (θ_t): $cos(6 \cdot [\theta_t - \varphi])$. That is, if we calculated the grid-angle correctly and the grid-angle is consistent among the data this regressor will mirror the neural data (**Figure 2.7H**). A higher value for this regressor thus indicates a better fit (**Figure 2.7I**).

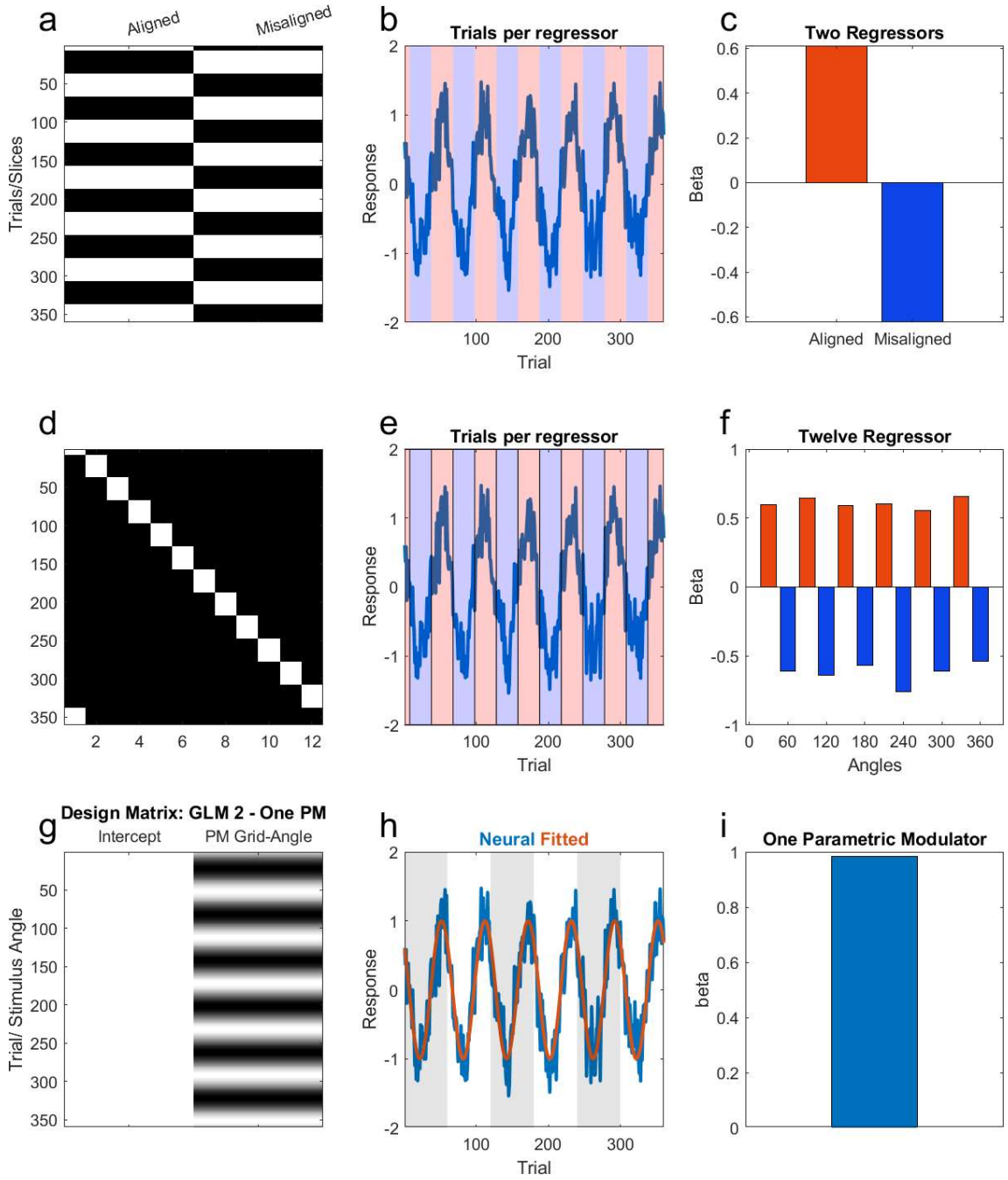


Figure 2.7. Results from GLM 2 for different regressors

There are a couple of standard analyses that one can do. Here we highlight three related but slightly different methods. **a-c**) Show the process of creating two regressors: one for the aligned trials and the other for the misaligned trials. **d-f**) We treat every possible movement direction for aligned and misaligned as a separate regressor. Resulting in 12 total regressors. **g-i**) Create one parametric modulator that fits the data based on the pattern it is supposed to show based on the grid-angle.

Resources

Github: https://github.com/kFrolichs/Thesis/tree/main/2.3-Grid-Code_Analysis

3. Study 1 - Finding social knowledge structures using computational models

3.1 Introduction and Hypotheses

In this first study we aimed to elucidate what strategies humans use when learning about others' personalities. Strategies are meant here as the largely unconscious processes at play during learning (Lockwood & Klein-Flügge, 2020). We use Rescorla-Wagner models (Rescorla & Wagner, 1972), that are a subset of general reinforcement learning (RL) models (Dayan & Niv, 2008; Dunne & O'Doherty, 2013; Joiner et al., 2017a; Niv & Langdon, 2016), as the foundation of the learning processes we aimed to unravel (see chapter 2.1 for a light introduction to these models).

However, it is clear that these models alone do not suffice in explaining any social learning process. Social learning is fast and intuitive, for example, when you see someone being mean you're quick to paint a picture of this person as not just being unkind but also as possessing other unfavorable traits. Sometimes we do not even need to see someone's behavior but just know that they are part of a specific group we have preconceptions about i.e., stereotypes (Kang et al., 2021; Mayer & Bower, 1986), to have an idea about their personality. The simple RW-rule cannot explain these more complex processes. Therefore we expanded on these models with knowledge structures. These knowledge structures are specific ways of representing information used during personality learning.

Research in personality established a multi-dimensional structure to human personality traits (Goldberg, 1990; R. McCrae & Costa Jr., 2008). Most commonly five independent dimensions are found. These so-called Big-5 factors (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness) can be used to describe people's personalities across the world (R. R. McCrae et al., 1999; Wiggins & Trapnell, 1997). Furthermore, people have been shown capable of abstracting from concrete examples to more general personality traits and the relations between them (Klein et al., 1992).

In this study, we want to find out if humans use these multi-dimensional structures during personality learning. We construct two knowledge structures that capture different granularities of these structures and add them to standard RW models (Figure 3.2A). This granularity reflected the

detail with which humans can represent others' personalities. First, coarse grained means participants only take the factors of the Big-5 into account i.e., they only keep track of 5 values, each for one factor of the Big-5. Second, fine granularity meant looking at trait words individually but also their relatedness with the other trait words. That is, participants not only update the current trait word they are learning about but also all the other trait words based on how related (i.e., correlated) they are to the current trait word. For example, "this person seems nice. I also expect them to be generous and kind (because these are related to nice) and not aggressive".

Another knowledge structure we wanted to capture was that of previously acquired knowledge or preconceptions. We named this knowledge structure Reference Point (RP), where the RPs are average expected values that one expects the person who is being learned about to possess. For example, "this person seems nice. I also expect them to be generous and kind and not aggressive".

We expect that both the granularity and the reference point knowledge structures underlie human personality learning. Where different granularities are used based on the circumstances in which the learners find themselves. For example, when information is sparse during learning, using coarse granular representation is a better use of resources than an unnecessary fine grained representation. In five different experiments we tested multiple variations of RW-models combined with the above mentioned knowledge structures. These models were compared to participant data to find what strategies might underlie human personality learning. Next to our main method of model comparison, we also conducted standard statistical analyses (e.g., regressions and correlations), model simulation and model validation to further support our findings.

3.2 Methods

3.2.1 Participants and task

In this study we conducted 5 experiments, each with the aim of elucidating a specific component of our computational models. All experiments were conducted in accordance with the declaration of Helsinki and approved by the local ethics committee (Ethik–Kommission der Ärtzekammer Hamburg, Number: PV5746). Participants gave written informed consent and were compensated the regular hourly fee. All experiments were of equal length taking approximately 30 minutes each. Experiments 1-3 were conducted on independent samples, whereas experiments 4 and 5 were tested on the same sample during the same session. Participants were only allowed to partake in one of the experiments 1-3 or the combination 4 and 5, and had to meet the following inclusion criteria: (1) age between 18 and 40 years, (2) German native speakers, (3) normal or corrected to normal vision, and (4) no history of neurological or psychiatric disorders. More participant details can be found in supplementary table 9.4. Experiments 4 and 5 were preregistered on the open science framework (OSF) (https://osf.io/8r6gv).

All experiments had the same task but differed in their content (trait words and profiles). In each experiment participants were asked to learn about 4 or 5 profiles, these profiles were either from independent self-ratings or constructed (see supplementary table 9.5). All participants were told was that the profiles were from a real person they had never met before. Learning happened on one profile at a time and profiles were counterbalanced between participants. For each profile participants saw 50 or 60 trials (depending on the experiment) in random order. Each trial started with a trait word (e.g., diligent), participants had four seconds to rate (on a scale of 1-8) what self-rating this particular profile had given themselves. After this rating phase they received immediate feedback about the actual self-rating for two seconds duration (Figure 1). This meant we expected participants to learn over time. After the learning task, participants were asked to give their own self-ratings on the same items as the ones used for the profiles they learned about.

As mentioned each experiment was slightly different from the previous to understand specific parts of each computational model. Experiment one was conducted on 4 real profiles, with 60 items in accordance with the Big-5. This experiment was used as a baseline for our models. In experiment two we wanted to test whether people would change their granularity if information got more sparse. Therefore we constructed profiles to have, on average, equal trait ratings but

without the similarities between items (i.e., the correlations originally found between items in behavioral data were absent from these profiles). Furthermore, only trait words from the factors agreeableness and conscientiousness were included. In experiment three, we again used real profiles but now, like experiment two, also using only trait words from the two Big-5 factors agreeableness and conscientiousness. In experiment four, we wanted to test the reference point in our models. To do so we changed the population of the profiles we tested from students to fashion models, expecting that participants would change their reference points to a more stereotypical one for the fashion models. Finally, in experiment five we wanted to test our models on a different set of items, to see if they were robust to this change. We therefore changed our stimuli from 60 trait words to 50 trait sentences taken from the German translation of the International Personality Item Pool (IPIP).

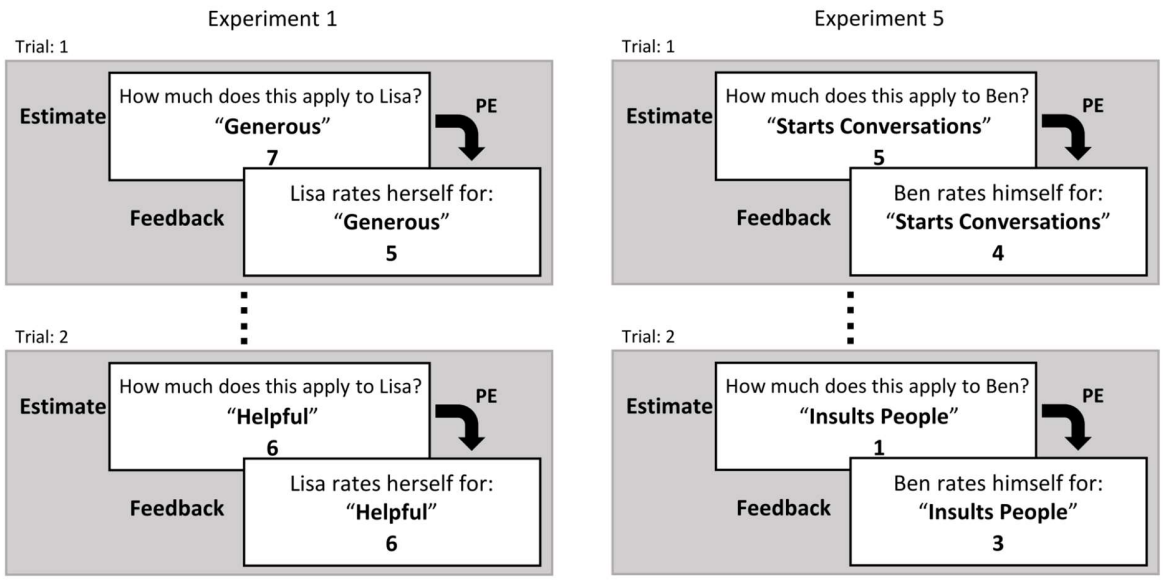


Figure 3.1. Overview of the experiment tasks.

In this study, we tested computational models of how humans learn about others. In five distinct experiments participants performed a social learning task on several profiles of other persons (with every profile being presented in a separate run). General overview of the learning task: Two trials for Experiment 1 (left) and Experiment 5 (right) are shown. During the learning task, participants estimated which self-ratings a person (here called profile) had given for specific traits (on a Likert scale from 1 does not apply at all to 8 does apply very much). After each estimate, participants received direct feedback in the form of the actual self-rating of that person. This process continued for all traits (in random order).

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.2.2 Statistical Analysis

We performed some standard statistical analyses to test whether participants were learning over time and what information they were using during learning. First, we calculated if there was a decrease of the average prediction error over time (i.e., an indication of learning). First, the average of the absolute PE per profile per trial was calculated. We then calculated the Pearson correlation coefficient on these averages together with the trial number to determine whether there was a decrease over time. That is, a negative correlation indicates an overall decrease of the PE over time. The same procedure was applied to real as well as simulated data.

Furthermore, we performed a GLM with three distinct regressors on the prediction errors. Each regressor represented a distinct property of our computational models. Briefly, these regressors were: 1) total number of previous trials seen for each item, (this assesses the relationship between the decrease of the PEs and the number of items seen previously and thus represents learning in the standard Rescorla-Wagner model), 2) total number of previous trials that are from the same factor as the current item (this assesses the relationship between the decrease of PEs and the number of items encountered from a specific factor and thus investigates the behavior captured by the coarse granularity models), 3) the summed absolute correlations of the previous items with the current item (this assesses the fine granularity models where the information content of all the previous items is weighted by their correlation to the current item). In a second-level analysis, participants' individual parameter estimates were subjected to a one-sided one-sample t-test to test if the slope was significantly different from zero in the negative direction (indicating a decrease of the absolute PE over time).

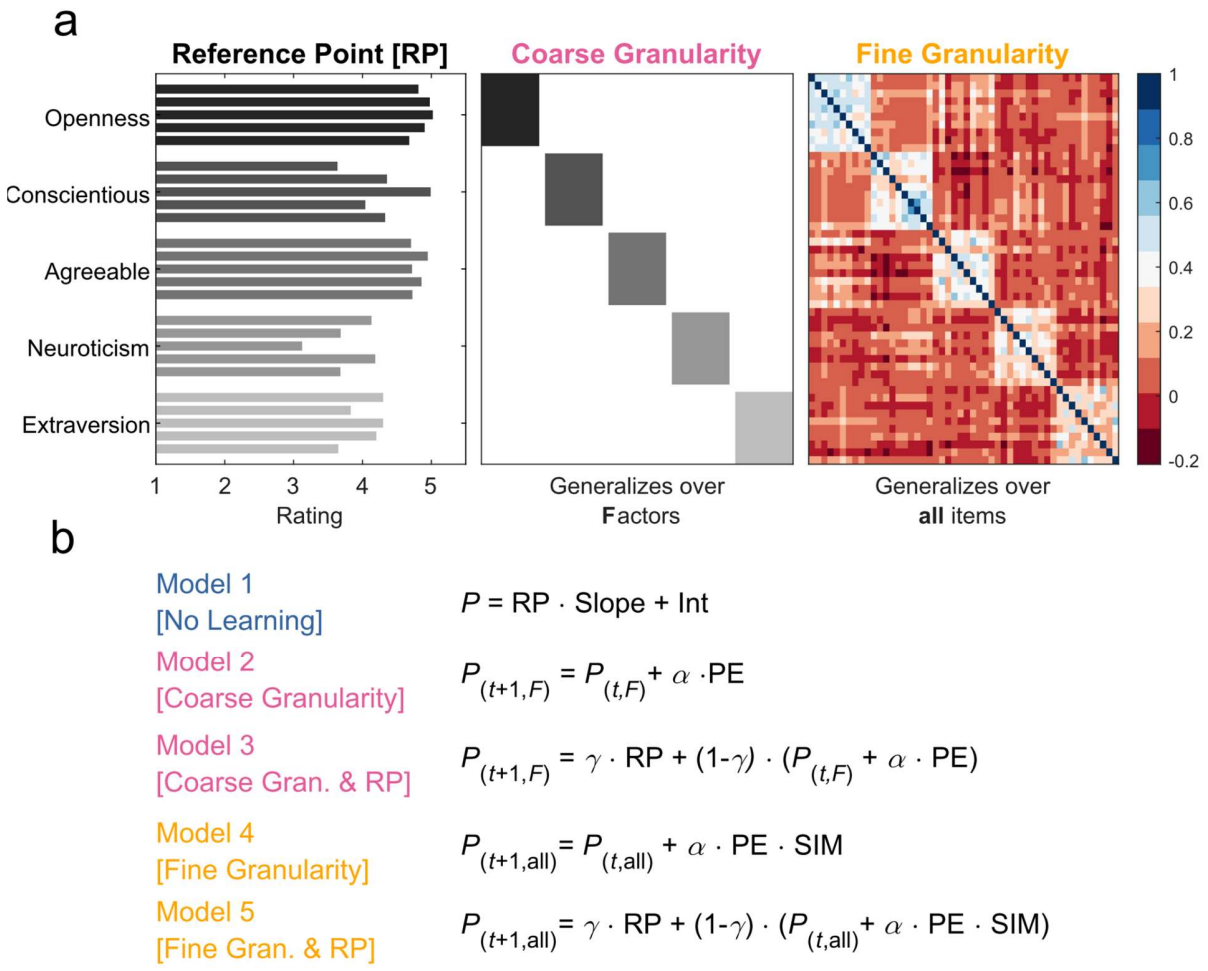


Figure 3.2. Overview of the models and knowledge structures

To explore participants' behavior we constructed five main computational models that made use of two main knowledge structures: Reference Points (RPs) and Granularity (G). **a**) The Reference Points represent what participants can use as a basis for estimating an average person (shown is a selection of student personality trait averages). Participants can use these RPs on each trait to compare this average rating with their current estimate for a specific person. Traits are ordered based on the Big-5 Factors (different shades of gray). Granularity (G) refers to the level of detail in the represented structure of others' personality traits. The granularity matrix generalizes the PEs across similar items in two distinct ways: for coarse granularity it generalizes per Big-5 factor, and for fine granularity it updates every individual trait based on how correlated they are to the current trait. **b**) Using both RPs and granularity the models can be divided into three sets, which are depicted in three different colors. First, No Learning (blue), consists of a single regression model, Model 1 [No Learning] that functions as a baseline model. Second, Coarse Granularity (pink), updates based on the (Big-5) factor to which the current adjective belongs. Model 2 [Coarse Granularity] uses the standard Rescorla–Wagner (RW) function to update the factor estimates and Model 3 [Coarse Granularity & Population RP] combines Model 2 with information from the RP. Third, Fine Granularity (orange), consists of two models that update all adjectives based on their correlation with the current trait. Model 4 [Fine Granularity] updates

all items according to the Fine Granularity and Model 5 [Fine Granularity & Population RP] combines model 4 with information from the RP (see Supplementary Fig. 1 for details on the models).
P prediction, Int intercept, RP reference point, α learning rate, PE prediction error, γ weighting parameter, F (generalizes over Factor) coarse granularity, All (generalizes over All items) fine granularity, SIM similarity matrix.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.2.3 Computational Models

We created several computational models based on the Rescorla-Wagner learning rule (see chapter 2.1 for a more detailed explanation together with code examples). These models made use of two knowledge structures, first, granularity represented the level of detail with which one represents the data (Figure 2A, middle and right). For this we constructed two levels of detail, coarse granularity, which only takes into account the levels of the Big-5 factors. That is, instead of tracking each trait word, it tracks the factors the trait word belongs to. Fine granularity, tracks each trait word separately and updates all trait words based on their correlation to the current trait word (e.g., if you learn someone is more generous than expected you will most likely also update how kind you expect them to be). The second knowledge structure was named reference point and represents prior knowledge one might have about (groups of) people (e.g., most people already have an idea of what the average student will be like and thus start with some expectations) (Figure 2A, left). We combined these knowledge structures with the RW-rule to create 5 distinct models.

Model 1, was created as a baseline model that represented No Learning. This was simply a linear regression on the reference points, where b_0 represents the intercept and b_1 the slope.

$$P = RP \times b_1 + b_0 \qquad (6)$$

Model 2, uses only the coarse granularity during learning. That is, it updates an average factor value for each of the 5 factors trait words can belong to. This model uses two free parameters, the first, α, is the learning rate [0-1] and the second, $starting\ value$, determined the value at which each factor was initialized [1-8] (the starting value cannot be found in the equation).

$$P_{(t+1,F)} = P_{(t,F)} + \alpha \times PE \qquad (7)$$

Model 3, combines the coarse granularity from model 2 with the reference points. That is, it updates its learning like model 2 but also adds information based on the reference points i.e., each item is updated based on the factor average but part of the trait word value is determined by the reference point. How much is relied on the learning from model 2 and how much based on the reference point is determined by the weighting parameter γ [0-1], Like model 2 it also uses the $learning\ rate$ and $starting\ value$.

$$P_{(t+1,F)} = \gamma \times RP + (1-\gamma) \times (P_{(t,F)} + \alpha \times PE) \qquad (8)$$

Model 4, uses only the fine granularity during learning. That is, it updates each trait word based on how correlated they are to the current trait word. The correlations between items (similarity) was calculated from a separate sample. This model has two free parameters, α, that determines the learning rate and $starting\ value$ that initializes all items at the beginning of learning.

$$P_{(t+1,All)} = P_{(t,All)} + \alpha \times PE \times SIM \qquad (9)$$

Model 5, combines the fine granularity from model 4 with the reference point, in a similar way to model 3. That is, it uses the weighting parameter γ to determine how much of model 4 is used and how much to rely on the reference point. Like model 4, this model also uses the free parameters α and $starting\ value$.

$$P_{(t+1,All)} = \gamma \times RP + (1-\gamma) \times (P_{(t,All)} + \alpha \times PE \times SIM) \qquad (10)$$

3.2.2.1 Additional Models

Our standard models use the average of a student population as the RP. For experiment 4 we added three additional models that use a stereotypic RP based on expected self-ratings from fashion models. These models are functionally the same as models 1, 3, and 5 except for the stereotypic RP (STE).

Model 1-STE: $P = STE \times b_1 + b_0$ (11)

Model 3-STE: $P_{(t+1,F)} = \gamma \times STE + (1-\gamma) \times (P_{(t,F)} + \alpha \times PE)$ (12)

Model 5-STE: $P_{(t+1,All)} = \gamma \times STE + (1-\gamma) \times (P_{(t,All)} + \alpha \times PE \times SIM)$ (13)

Similar to the use of the stereotypic RP we created 3 models that use participants' self-ratings as the RP ($SELF$).

Model 1-SELF: $P = SELF \times b_1 + b_0$ (14)

Model 3-SELF: $P_{(t+1,F)} = \gamma \times SELF + (1-\gamma) \times (P_{(t,F)} + \alpha \times PE)$ (15)

Model 5-SELF: $P_{(t+1,All)} = \gamma \times SELF + (1-\gamma) \times (P_{(t,All)} + \alpha \times PE \times SIM)$ (16)

To explore whether participants learned differently from positive and negative feedback we created models that use two free parameters for the learning rate. One for positive feedback (i.e., where the participants' estimate was too low) and the other for negative feedback (i.e., where the participants' estimate was too high). All models, except for model 1 (which does not update via the learning rate), were adapted to use these two learning rates.

$\alpha = \alpha^{-} if\ PE < 0$ (17)

$\alpha = \alpha^{+} if\ PE \geq 0$ (18)

3.2.4 Model Fitting and Comparison

Model fit was determined with the sum of squared errors (SSE), all models were initialized with their free parameters at the midpoint of their respective boundaries. Optimal parameters were searched using the non-linear Nelder-Mead simplex search algorithm (implemented as *fminsearch* in MATLAB), to minimize the SSE. To fairly compare models based on the lowest SSE achieved we used the Bayesian Information Criterion (BIC). This BIC punishes models for their complexity (i.e., number of free parameters). Because more complex models can achieve better fits purely through this complexity.

$SSE = \sum_{i=1}^{n} (Outcome_i - Prediction_i)^2$ (19)

$BIC = n \times ln(\frac{SSE}{n}) + k \times ln(n)$ (20)

Where n is the number of trials and k the number of free parameters. Models were compared using fixed-effect analysis using the log-group Bayes factor. That is, the BIC scores across all participants are summed for each model separately and then subtracted from the worst scoring model, the best

model has the lowest score and is thus chosen as best at explaining the data it was fit to. For random-effects analysis we calculated the posterior exceedance probability. This comparison was performed using the Bayesian Model Selection (BMS) procedure implemented in the MATLAB toolbox SPM12 (http://www.fil.ion.ucl.ac.uk/spm/; *spm_BMS*).

Best Performing Models

Regular model fitting procedures fit the models to participants' data. This provides information on participants' behavior, that is, the models give us an insight into the strategies that the participants are using. However, we can also look at the models themselves and find which of these strategies is best for the task at hand. We do this by fitting the models on the same task as the participants and then evaluate how well they perform. Ideally, this best performing model is the same as the model that fits participants data best, indicating that participants use the best strategy (out of those that we test).

3.2.5 Model Checks

Confusion Matrices

Model comparison wants to compare models that are different from each other. That is, exhibit different behavior when fit on the data. This allows us to determine what strategies participants likely use. This is all performed around the assumption that models are different from each other, but also exhibit different behavior. The confusion matrix tests this assumption. To do so, we first simulate data with each of the models and subsequently fit all models on this simulated data. Ideally, the models that were used to create the data will also fit that data best. If this is not the case (i.e., there is confusion) we have to take models out of our sample or adapt them to avoid confusion. We calculated a confusion matrix for every experiment separately. First we randomly drew 200 parameter values for each model to simulate data with. The parameter values were always between the values [0.2 0.8]. These values were chosen because the complex models were based on simpler models who's use within these models depended on a weighting parameter. A weighting parameter at one of the extremes [0 or 1] would mean a simpler model would be used, thus causing confusion. After creating the simulated data we added random normally distributed noise and all models were fit on this simulated data. Ideally the model that simulated data is also the best fitting model thus resulting in an identity matrix.

Parameter Recovery

One of the goals of the model fitting procedure is to find the parameter values coupled to the strategies (i.e., models) from the participants' data. To make sure that the parameter values we find in our modeling procedure are also the ones used to create this data, we use parameter recovery. We therefore simulate data with known parameters and recover these with the normal fitting procedure. Ideally, the fitted parameters (i.e., output) match the inputted parameters, this would mean the parameters get recovered accurately.

In a similar vein to the confusion matrix, we simulate data 200 times with randomly sampled parameters between [0 1] and add noise drawn from the normal distribution. After this we fit all models on their simulated data and compare the actual parameters with the retrieved parameters. In an ideal situation one would retrieve the parameters with a perfect correlation to the ones used to simulate the data.

3.3 Results

Results from all five experiments will be described separately. Each experiment built on the previous.

3.3.1 Experiment One

In the first experiment we wanted to set a baseline for the other experiments. We therefore tested our models on regular data i.e., 60 trait words spread over five factors, for four different profiles that were based on real self-ratings. Using standard statistics we found that participants learned over time. That is, a Pearson correlation of the average absolute prediction error and the trial number was negative indicating learning ($r(58) = -0.523$, $p < 0.001$). The same analysis on simulated data indicated similar learning ($r(58) = -0.563$, $p < 0.001$) (Figure 3C).

Further evidence of learning came from a GLM with three regressors each aimed at a specific part of our models. For this GLM, the third regressor (summed absolute correlations) was significant ($t(34) = -1.7109$, $p = 0.0481$), showing participants' use of similarities between traits.

Results from model comparison showed that model 5 (Fine Granularity and Population RP), was the winning mode (Figure 3A)l. Further cementing our belief that participants use trait similarities together with the RP during learning. Model comparison of simulated data showed model 5 to be the best performing model i.e., out of the 5 possible models this was the most optimal model to use for this task (Figure 3B).

Having gathered some evidence for participants' use of the fine granularity, we hypothesized they would switch to coarser granularity when presented with less informative profiles. This was investigated in experiment two.

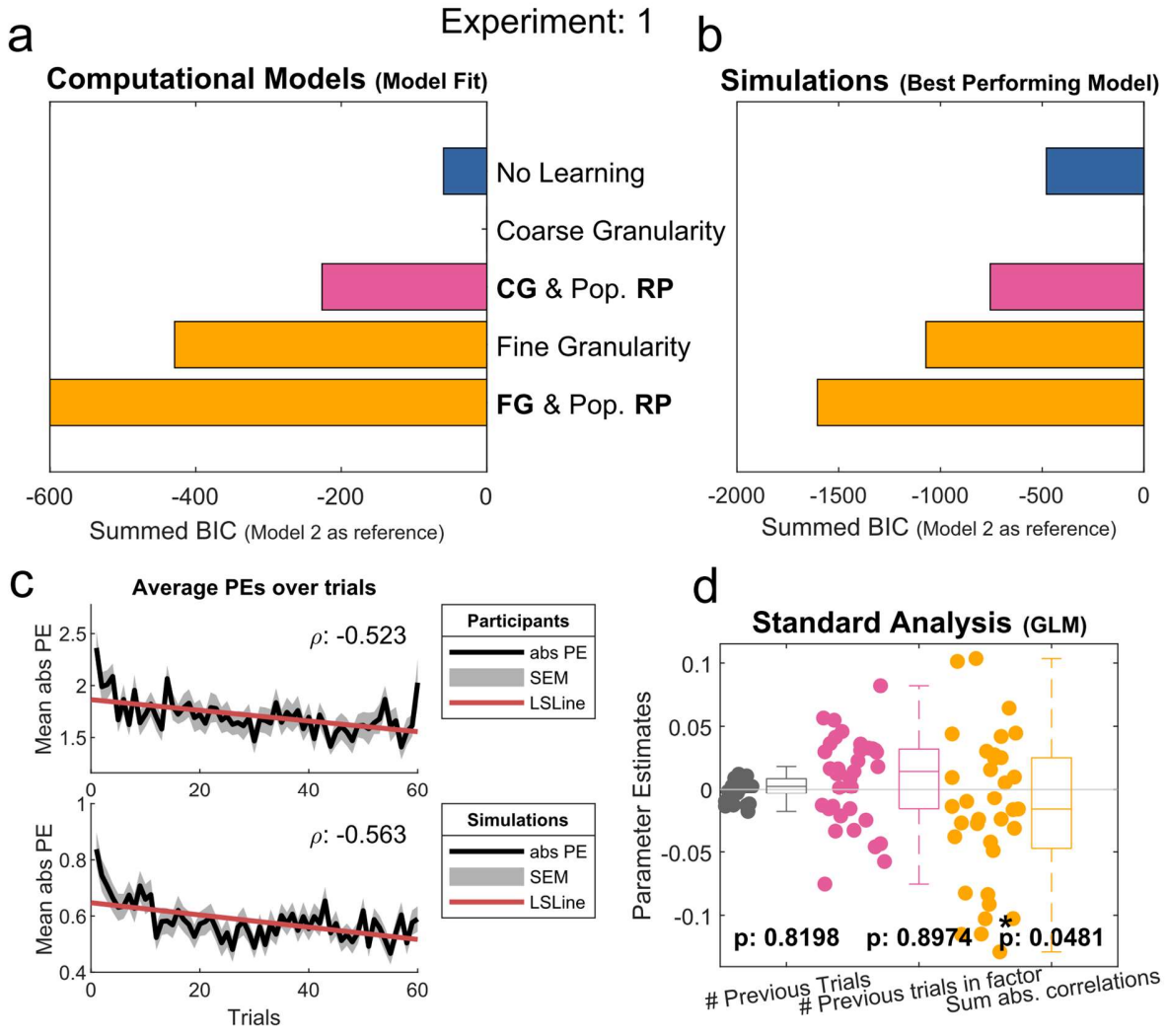


Figure 3.3. Overview of the main analyses for experiment 1.

In this experiment, participants learned about the personalities of four strangers. Results indicate that participants used fine-grained correlation structures during learning. **a**) Model comparison results using fixed-effects analysis (losing model as reference) indicate Model 5 [Fine Granularity (FG) & Population Reference Point (RP)] as the best fitting model ($n = 35$). **b**) Simulated data ($n = 35$), the best performing model indicates which of the models performs the task most optimally. The best performing model (Model 5) demonstrates that participants used the best strategy. **c**) A decrease of the prediction errors (PEs) over time can be interpreted as learning. Both plots display the average absolute PEs over time ± SEM. We calculated a pairwise Pearson correlation between trial number and the mean absolute PEs to determine if the PEs decrease over time. Top) Participants' data shows a decrease in the PEs over time (ρ:−0.523, least squares line (red)). Bottom) Simulated data using Model 5 shows a similar decrease in PEs over time (ρ: −0.563). **d**) General linear model (GLM) on three core model features: (1) Rescorla–Wagner RL, 2) coarse models, and 3) fine models that predict the accuracy per trial per participant. Only the third regressors was significant ($n = 35$), indicating participants' use of fine granularity: (one-sided t-test) regressor 1: $t(34) = 0.927$, $p = 0.8198$, regressor 2: $t(34) = 1.2915$, $p = 0.8974$, regressor 3: $t(34) = -1.7109$, $p = 0.0481$.

Individual data points are participants' parameter estimates which are summarized by boxplots (median (middle line), 25th, and 75th percentile (box), the whiskers extend to most extreme data points not considered outliers (1.5 times interquartile range), outliers are indicated with + signs). Conclusions based on this GLM should take into account that all three regressors are highly correlated (ρ between 0.76 and 0.92). [One-sided *t*-test; * indicates $p < 0.05$, ** indicates $p < 0.001$, no correction for multiple comparisons]. CG coarse granularity, FG fine granularity, RP reference point, # number of, PEs prediction errors, SEM standard error of the mean, LSLine least squares line.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.2 Experiment Two

In the second experiment we presented participants with constructed profiles. That is we kept everything the same but instead of using profiles with real self-ratings we created them. Importantly, these constructed ratings still had the same mean and standard deviation as the real factor ratings but the similarities inherent in real ratings were absent. Like experiment one, participants were asked to rate four profiles on 60 trait words, however the number of factors was decreased to two (agreeableness and conscientiousness).

A Pearson correlation coefficient of the absolute PE and the trial number was negative, this decrease over time indicated learning $r(58) = -0.564$, $p<0.001$ (Figure 4C). Model simulations also showed a decrease over time ($r(58) = -0.489$, $p < 0.001$).

Next we calculated a GLM with three regressors that showed participants' use of both coarse and fine granularities as well as learning over trials (Figure 4D).

Results from model comparison confirmed our hypothesis that participants would use a coarser granularity during learning i.e., Model 3 [Coarse Granularity and Population RP] was the best fitting model for both fixed- and random-effects analyses. Model simulations found Model 2 [Coarse Granularity] to be the best performing model, indicating that participants could have improved their performance if they would not have used the reference point.

In this experiment we not only changed the profiles from real self-ratings to constructed but we also used less factors. To make sure this narrower item set was not the cause of participants' change in strategy we conducted a third experiment with real self-ratings on two factors.

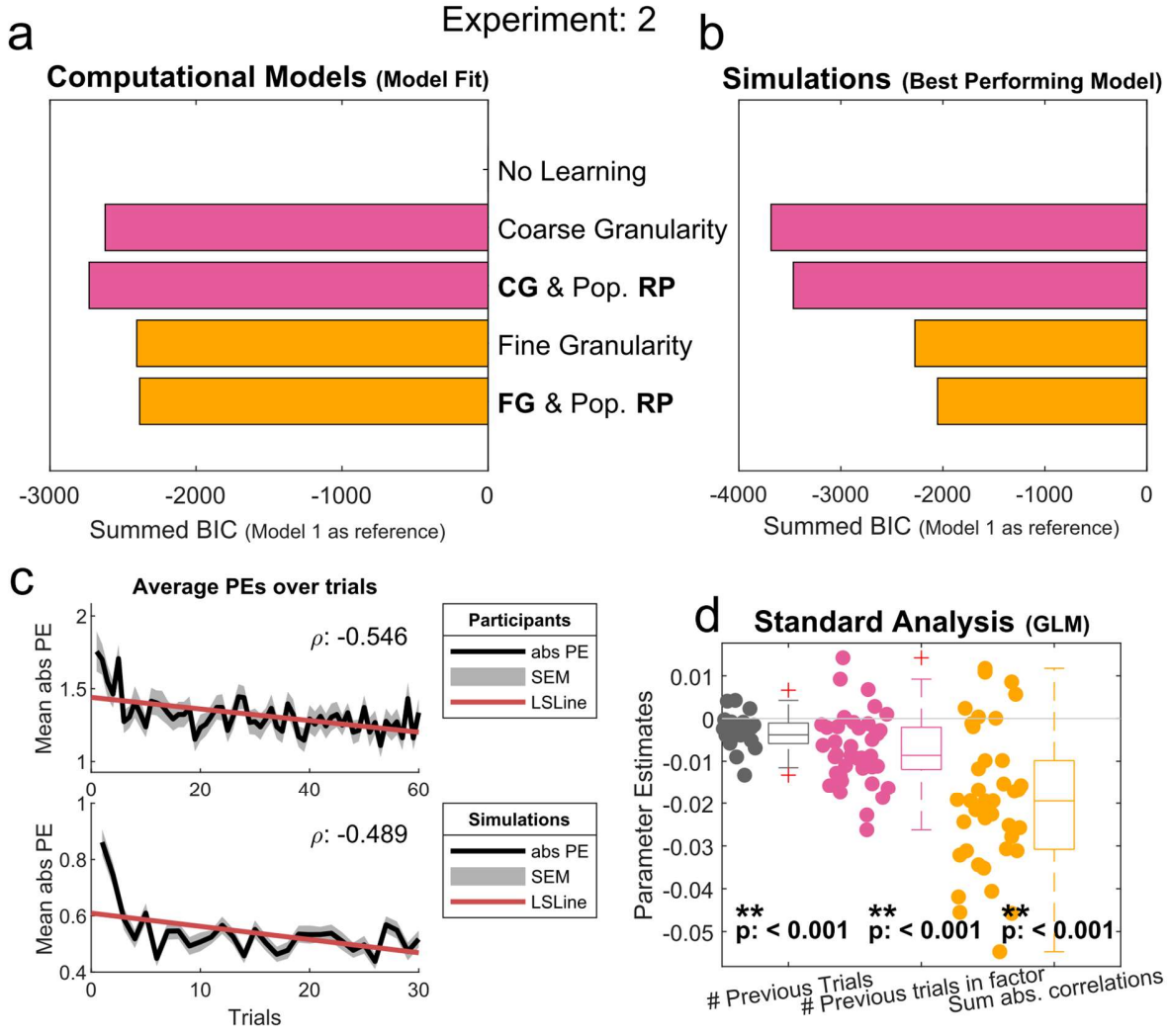


Figure 3.4. Overview of the main analyses for experiment 2.

In this experiment participants learned about artificial profiles which did not have the trait similarity structures. Results indicate that participants use a coarse granularity structure when less social information is present. **a**) Model 3 [Coarse Granularity & Population RP] is the best fitting model ($n = 41$). This model uses the average population as a reference point and coarse granularity for generalization. **b**) Simulated data for the best performing model ($n = 41$). Unlike participants' data, Model 2 [Coarse Granularity] was the best performing model, demonstrating that participants could have used a more optimal strategy. **c**) Both plots display the average absolute PEs over time ± SEM. Top) Participants' data shows a decrease in the PEs over time (ρ:−0.546, least squares line, red), this indicates participants were learning over time. Bottom) simulated data from the best fitting model (Model 3) shows a similar decrease in PEs over time, indicating that the models learned in a similar way to participants. **d**) All three regressors (representing: 1 RW learning, 2 Coarse granularity, 3 Fine granularity), were significant (one-sided t-test), regressor 1: $t(40) = -5.4617$, $p < 0.001$, regressor 2: $t(40) = -5.7377$, $p < 0.001$, regressor 3: $t(40) = -7.7059$, $p < 0.001$, indicating that participants ($n = 41$) learned over time but also made use of both coarse and fine granularity. However, these regressors were correlated and conclusions regarding this GLM should thus be drawn with caution. Participants' parameter estimates (for each

regressor) are indicating by the individual data points, which are summarized by the adjacent boxplots of the same color.

The boxplots indicate the median (middle line), and the box is formed by the 25th, and 75th percentile. The whiskers extend to most extreme data points not considered outliers (1.5 times interquartile range), outliers are indicated with + signs. [One-sided *t*-test; * indicates $p < 0.05$, ** indicates $p < 0.001$, no correction for multiple comparisons]. CG coarse granularity, FG fine granularity, RP reference point, # number of, PEs prediction errors, SEM standard error of the mean, LSLine least squares line.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.3 Experiment Three

Experiment 3 combined parts from both previous experiments, like experiment one, it used real self-ratings. However, like experiment two, these items were on only two factors (agreeableness and conscientiousness). Like both experiments each of the four profiles consisted of 60 trait words (30 for each factor). We hypothesized that with these real self-ratings participants would once again use the fine granularity.

First, we calculated the Pearson correlation coefficient for the trial number and absolute PE. This revealed a negative correlation ($r(58)=-0.42, p<0.001$, Figure 5C top) which indicates participants learned over time. Similarly, the Pearson correlation for trial number and the absolute PE for simulated data was negative ($r(58) = -0.764$, $p < 0.001$, Figure 5C bottom). This indicates the simulations captured some of this learning.

Next, a GLM with three regressors that each capture an important part of our models found all three regressors to be significant, showing participants learned over time but also used both coarse and fine granularity during learning (Figure 5D).

Model comparison confirmed our hypothesis, both fixed- and random-effect analyses showed Model 5 [Fine Granularity & Population RP] to be the best fitting model (Figure 5A). Model simulations showed that Model 5 was also the best performing model of the models in our set (Figure 5B).

The first three experiments all found evidence for the flexible use of granularity based on information from the environment. Having found this we next wanted to test whether participants would also change their reference point. To do this we changed the profiles from students to fashion

models, expecting participants to perceive fashion models as having different personalities from students.

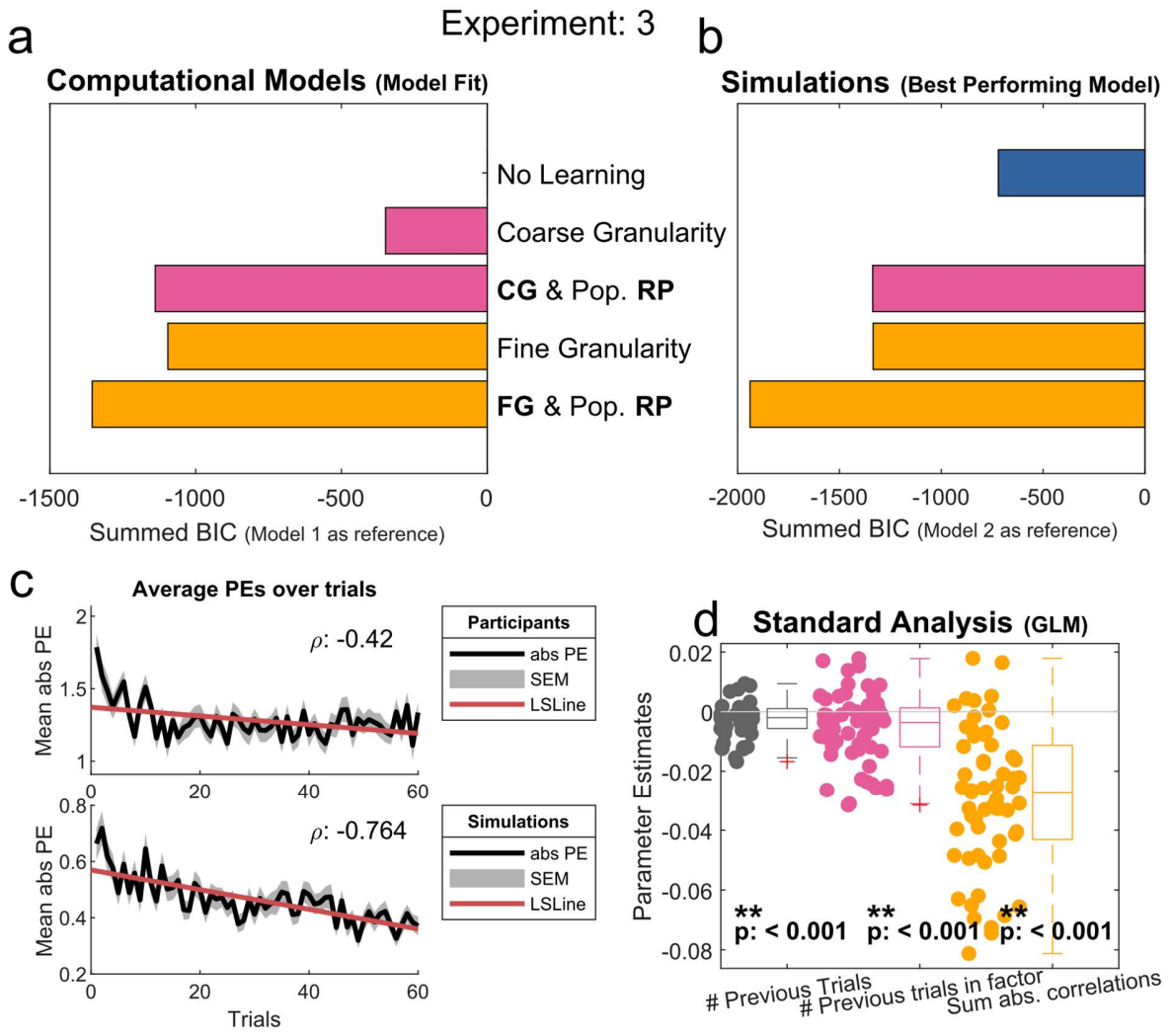


Figure 3.5. Overview of the main analyses for experiment 3.

In this experiment participants learned about real profiles on two Big-5 factors. Results indicate participants' use of a fine correlation structure. **a**) Similar to experiment 1, Model 5 [Fine Granularity & Population RP] is the best fitting model for experiment 3 ($n = 59$). This model uses the average population as a reference point and fine granularity for generalization. **b**) Simulated data to find the best performing model ($n = 59$). In line with results from the participants, model 5 [Fine Granularity & Population RP] was the best performing model, demonstrating that participants used the best possible strategy. **c**) Both plots display the average absolute PEs over time ± SEM. Top) Participants' data shows a decrease in the PEs over time (ρ:−.42, least squares line (red)), which indicates that participants learned over time. Bottom) Simulated data from the best fitting model (Model 5) also shows a decrease in PEs over time (ρ:−0.764), showing that the models learned in a similar way to participants. **d**) All three regressors (representing: 1 RW learning,

2 Coarse granularity, 3 Fine granularity), were significant (one-sided *t*-test), regressor 1: $t(58) = -3.414$, $p < 0.001$, regressor 2: $t(58) = -3.6269$, $p < 0.001$, regressor 3: $t(58) = -9.4348$, $p < 0.001$, showing participants ($n = 59$) learned over time but also made use of both coarse and fine granularity.

Individual parameter estimates are indicated by the colored dots, which are summarized by the adjacent boxplots (median (middle line), 25th, and 75th percentile (box), most extreme points not considered outliers (whiskers), outliers (1.5 times interquartile range) indicated with + signs). Due to high correlations between these regressors, conclusions regarding these regressors should be drawn with caution. [One-sided *t*-test; * indicates $p < 0.05$, ** indicates $p < 0.001$, no correction for multiple comparisons]. CG coarse granularity, FG fine granularity, RP reference point, # number of, PEs prediction errors, SEM standard error of the mean, LSLine least squares line.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.4 Experiment Four

In experiment four we changed the profiles from a student population to fashion models. We expected participants to perceive fashion models' personalities as different from students and to accommodate for this change by using a different (stereotypical) RP. To test for the stereotypical views we added three computational models to our existing set of five. These additional models used a stereotypical RP (i.e., indicated by the -STE extension) which we gathered by asking participants to rate the expected fashion model self-ratings before the start of the experiment. We conducted a paired sample t-test on the average student (M: 6.01, SD: .69) and fashion model (M: 5.13, SD: .32) reference points which indicated a significant difference between them $t(59) = -9.7137$, $p<0.001$.

First, to investigate whether there was learning, we calculated a Pearson correlation coefficient between the trial number and the absolute PE. This produced a negative correlation $r(58) = -0.307$, $p<0.001$, which indicated learning over time (Figure 6C, top). The same analysis for simulated data also found indications for learning ($r(58) = -0.788$, $p < 0.001$, Figure 6C, bottom).

A GLM with three regressors, each representing a significant part of our models, showed participants' use of fine granularity (Figure 6D). That is, the third regressor which contained the sum of all previous correlations, was significantly different from zero ($t(28) = -2.6794$, $p = 0.0061$).

Model comparison confirmed our hypothesis: Model 5-STE [Fine Granularity & Stereotype RP] was the winning model for both fixed- and random-effects analysis. However, model simulations showed Model 1 [No Learning] to be the best performing model. Indicating a discrepancy between the best strategy and those used by participants.

These final results concluded evidence for our models usefulness in explaining the use of both granularity as well as the reference points. In our final experiment we wanted to explore whether different stimuli could be used successfully as well.

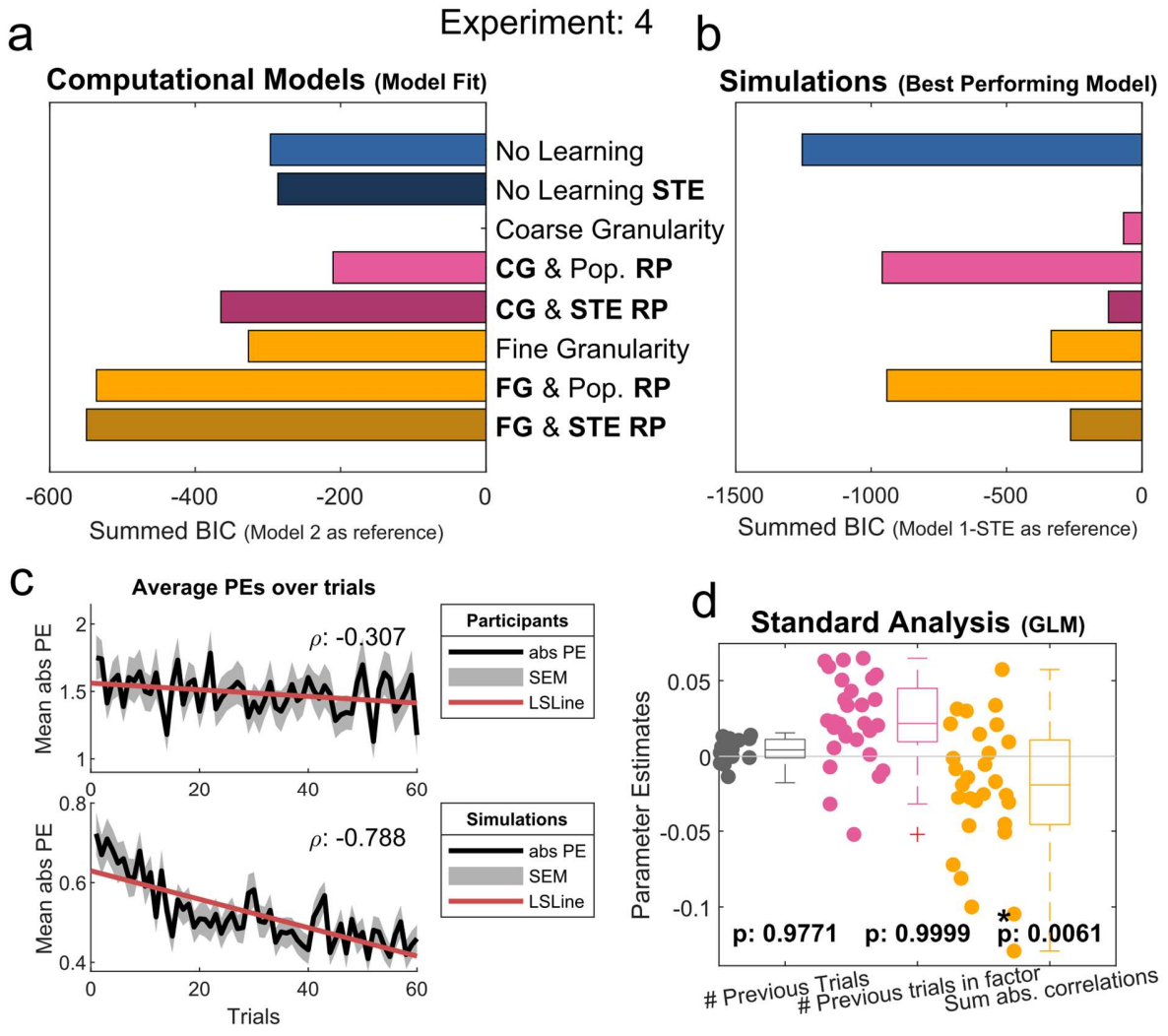


Figure 3.6. Overview of the main analyses for experiment 4.

In this experiment participants learned about an out-group (i.e., fashion models) instead of the regular in-group (i.e., students). Additionally, three models were added to our original set of five models to capture stereotypic inclinations (STE). These stereotypic models have a darker color in the figures and are indicated by STE in the model names. **a)**

As hypothesized, the best fitting model was Model 5-STE [Fine Granularity & Stereotypic RP] (n = 29). This model uses the expected stereotypical self-ratings from models as a reference point and fine granularity for generalization. **b**) Simulated data for the best performing model (n = 29). Contrary to participants' data the best performing model was model 1 [No Learning]. This indicates that participants used too complex a strategy for learning about the fashion models. **c**) Both plots display the average absolute PEs over time ± SEM. Top) Participants' data show a decrease in the PEs over time (ρ:−0.307, least squares line (red)), an indication of learning over time. Bottom) Simulated data from the best fitting model (Model 5-STE) show a large decrease in PEs over time (ρ:−0.788). **d**) Of the three regressors (representing: 1 RW learning, 2 Coarse granularity, 3 Fine granularity), only the third regressor was significant (one-sided t-test), regressor 1: $t(28) = 2.0906$, $p = 0.9771$, regressor 2: $t(28) = 4.3546$, $p = 0.9999$, regressor 3: $t(28) = -2.6794$, $p = 0.0061$, indicating participants (n = 29) used fine granular representations during learning.
Individual data points represent participants' parameter estimates. Boxplots summarize these parameter estimates (median (middle line), 25th, and 75th percentile (box), most extreme points not considered outliers (whiskers), outliers (1.5 times interquartile range) indicated with + signs). Due to the high correlations between regressors one should be careful when drawing conclusions based on these regressors. [One-sided t-test; * indicates $p < 0.05$, ** indicates $p < 0.001$, no correction for multiple comparisons]. CG coarse granularity, FG fine granularity, RP reference point, # number of, PEs prediction errors, SEM standard error of the mean, LSLine least squares line.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.5 Experiment Five

Having found evidence for the use of both levels of granularity as well as different reference points based on the information in the task we next wanted to test our modeling paradigm on a different set of trait items. This new set of trait items consisted of 50 items from the German translation of the International Personality Item Pool (IPIP). These 50 items were equally divided over the 5 factors from the Big-5, and consisted of sentences such as "I am the life of the party" and "I sympathize with others' feelings". We expected our models to function just as well on the learning data from these traits as they did on the other trait data.

First we calculated a Pearson correlation coefficient between the trial number and the absolute PE. This resulted in a negative correlation $r(48) = -0.311$, $p < 0.001$ (Figure 7C, top), showing participants learned over time. This same analysis on simulated data also resulted in a negative correlation ($r(48) = -868$, $p < 0.001$, Figure 7C, bottom).

Further, a GLM with three regressors that each represented a part of our models, did not find any evidence of participants using these knowledge structures.

Both fixed- and random-effects analyses found model 4 [Fine Granularity] to perform best. This indicated that participants used fine granular representations but not the reference point during learning. Model simulations corroborated with these findings and indicated that model 4 [Fine Granularity] was indeed the best model to use for this task.

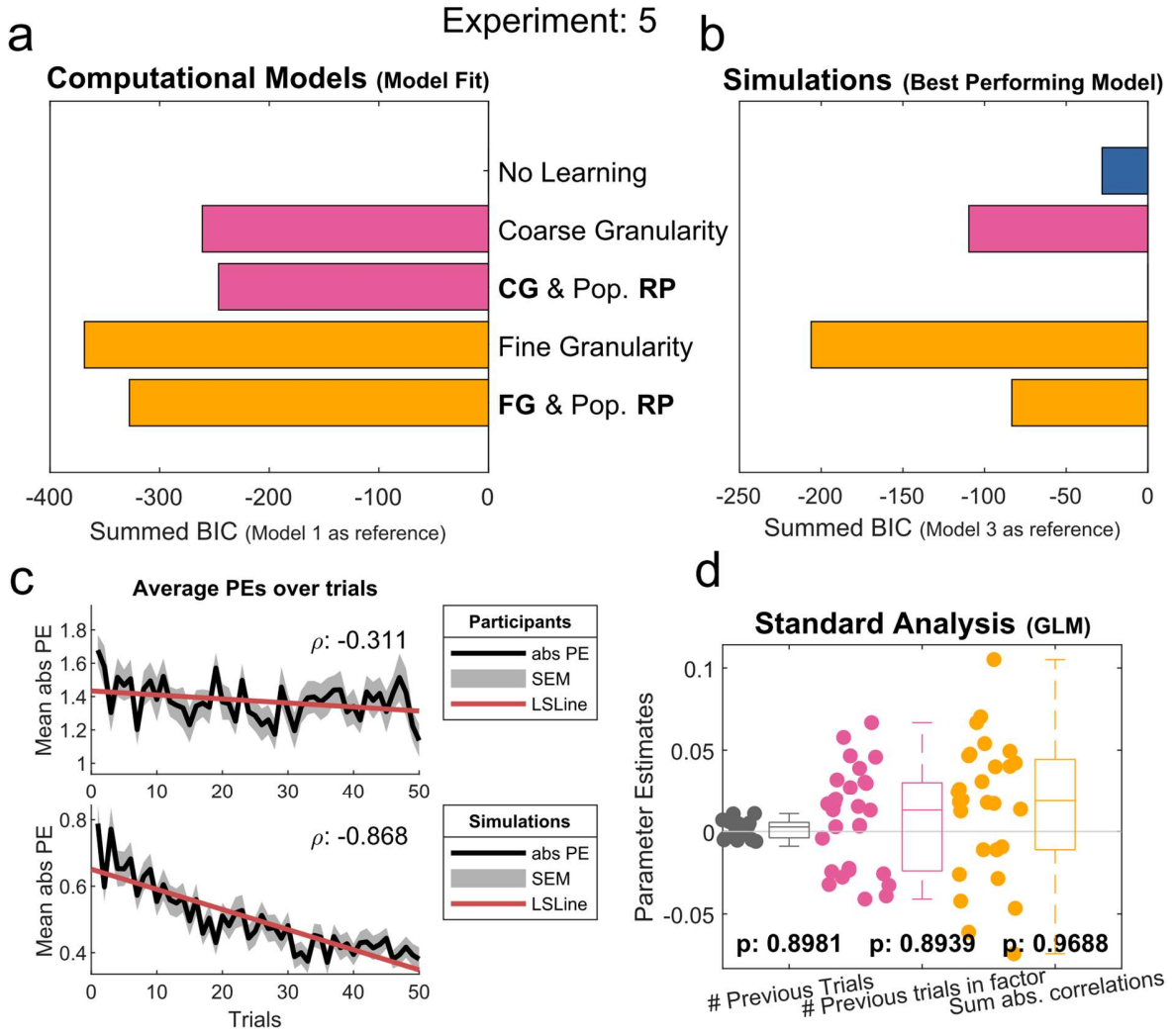


Figure 3.7. Overview of the main analyses for experiment 5.

In this experiment participants learned about the German translation of the 50-item International Personality Item Pool (IPIP). **a**) Model 4 [Fine Granularity] was the best fitting model ($n = 28$). **b**) Model simulations ($n = 28$) confirmed that Model 4 was the best performing model and thus the best strategy to use for this specific experiment. Showing that the Population RP did not add enough information to be useful. **c**) Both plots display the average absolute PEs over time ± SEM. (Top) Participants' data shows a decrease in the PEs over time (ρ:−0.311, least squares line (red)), indicating participants learned over time. Bottom) Simulated data from the best fitting model (Model 4) shows a large decrease in PEs over time (ρ:−0.868). Replicating participants' learning during the task. **d**) A GLM with three

regressors (1 RW learning, 2 Coarse granularity, 3 Fine granularity) resulted in no significant regressors ($n = 28$, one-sided t-test) regressor 1: $t(27) = 1.3021$, $p = 0.8981$, regressor 2: $t(27) = 1.2775$, $p = 0.8939$, regressor 3: $t(27) = 1.9446$, $p = 0.9688$.

Participants' individual parameter estimates are indicated by colored data points. These parameter estimates are summarized by the boxplots of the same color. Boxplots indicate the median (middle line), and the box by the 25th, and 75th percentile, the whiskers are the most extreme points that are not considered outliers (1.5 times interquartile range), outliers are indicated with + signs. [One-sided t-test; * indicates $p < 0.05$, ** indicates $p < 0.001$, no correction for multiple comparisons]. CG coarse granularity, FG fine granularity, RP reference point, # number of, PEs prediction errors, SEM standard error of the mean, LSLine least squares line.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.6 Model Robustness and Distinguishability

For every experiment separately we performed analyses to determine if models were both robust and distinguishable. That is, we calculated a confusion matrix and parameter recovery for all models for all experiments.

3.3.6.1 Confusion Matrix

The confusion matrices for all experiments were satisfactory: Experiment 1 all models >96.5% (Figure 8A), Experiment 2 all models > 99% (Figure 8B), Experiment 3 all models > 98% (Figure 8C), Experiment 4 all models > 97.5% (Figure 8D), Experiment 5 all models > 92.5% (Figure 8D). Indicating that for each experiment our models could be distinguished from each other.

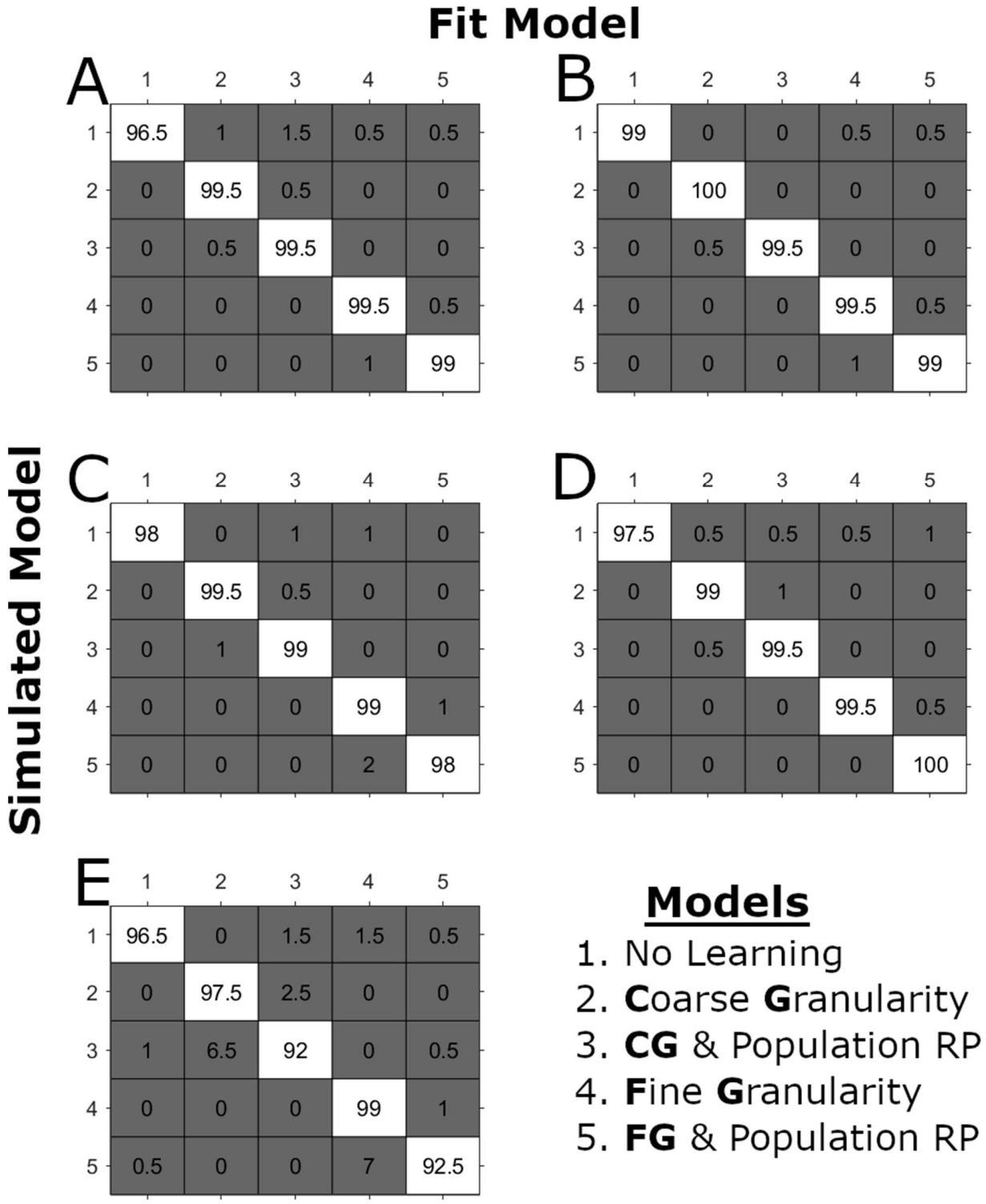


Figure 3.8. Confusion matrices for every experiment.

To test whether data created with a model would also be best fitted by that model, a confusion matrix was created for every experiment separately by randomly sampling parameter values 200 times. With these random parameter values, data was simulated and noise, sampled from a standard normal distribution, was added. After this, the models were

fitted to the simulated data. Ideal recovery is reflected by the identity matrix (i.e., all values on the diagonal being 100%). Model recovery for all experiments indicated that our models were recovered correctly in most cases (>92%), i.e., models were distinguishable. A-E represent experiments 1-5.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.6.2 Parameter Recovery

The parameter recoveries were performed on each of the model's free parameters. Across all 5 experiments and all models the correlation coefficient for the slope/ learning rate was .74 at its lowest (Figure 10). The weighting factor, which was only used in models 3 and 5, never had a lower correlation coefficient than .89 (Figure 9). Finally, the lowest correlation coefficient for intercept/ starting value was .51 (Figure 11). This showed that all parameters could be recovered satisfactorily.

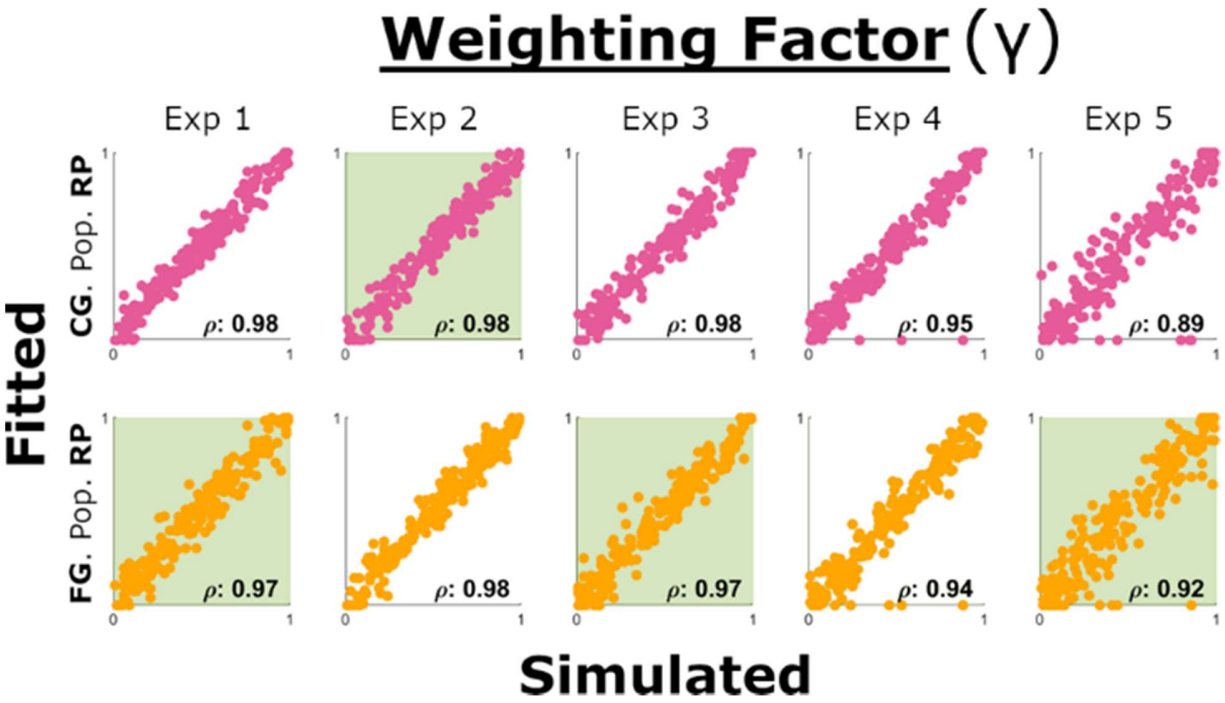


Figure 3.9. Parameter recovery for gamma (weighting factor).

To check whether model fitting returns meaningful parameter values, we performed parameter recovery on all models and all experiments separately. We simulated data using random parameter values, to which noise was added. After this, all models were fitted to this simulated data and parameters recovered. We measured recovery by calculating the Pearson's correlation coefficient (ρ) where a value closer to 1 indicated better recovery. The Gamma parameter was

only used in Model 3 [Coarse Granularity and Population Reference Point] and Model 5 [Fine Granularity and Population Reference Point]. Models are ordered along the rows and experiments on the columns. Recovery for all models and all experiments is good across the whole range of values. Green boxes indicate when a model was the best fit for that specific experiment.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

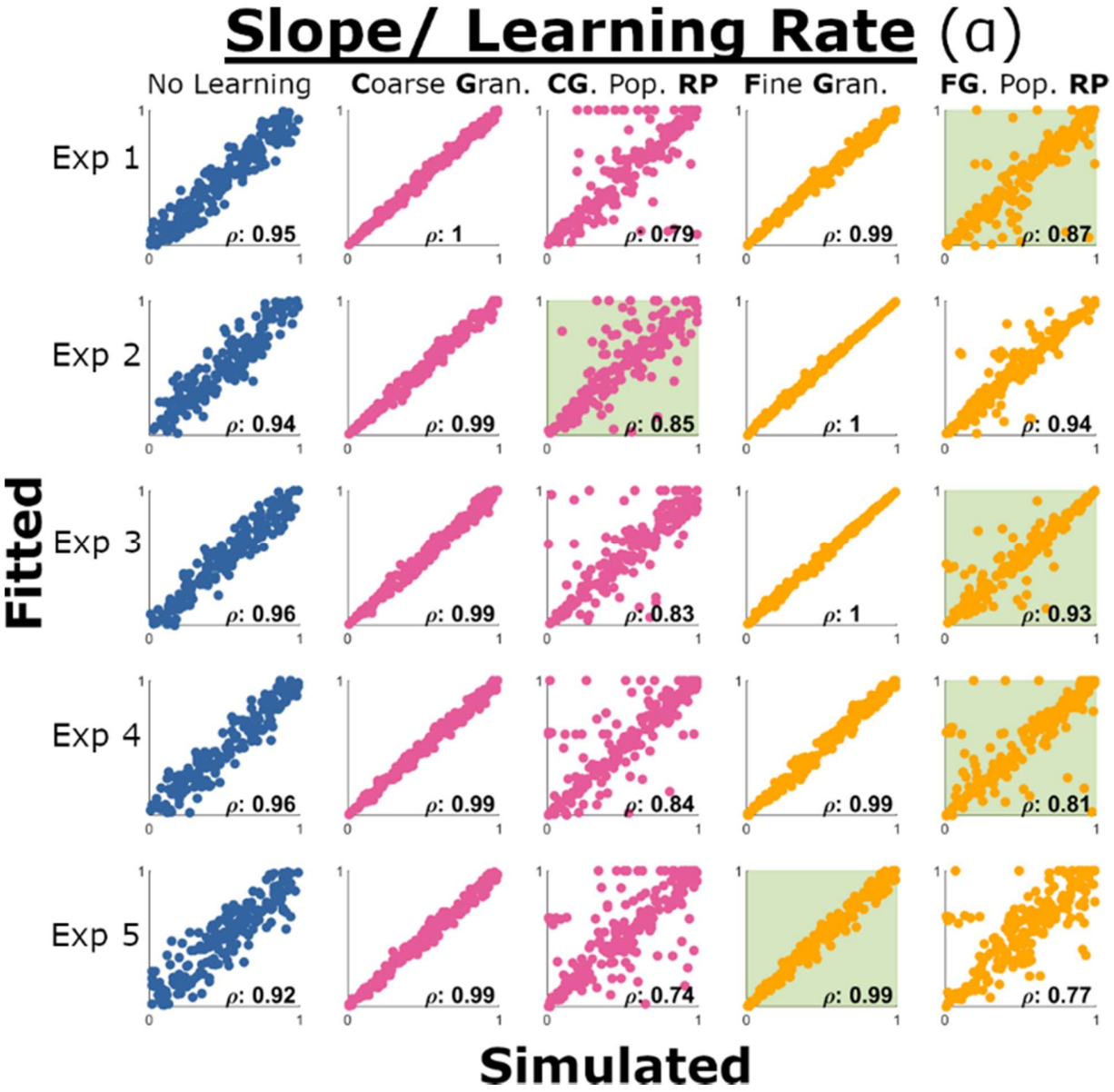


Figure 3.10. Parameter recovery for alpha (learning rate).

See Figure 9, for an explanation of how we conducted the parameter recovery. Models are ordered along the columns, every row is a separate experiment. For Model 1 (first column, blue dots) we looked at the regression slope and for all the other experiments at the learning rate (α). Parameter recovery for all experiments and all models was well within acceptable bounds. Green boxes indicate the winning model for that specific experiment. For Experiment 4, we only added the standard 5 models, choosing to omit the stereotype models. This was done because the stereotype models are functionally the same when it comes to simulated data and thus would not add any information.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

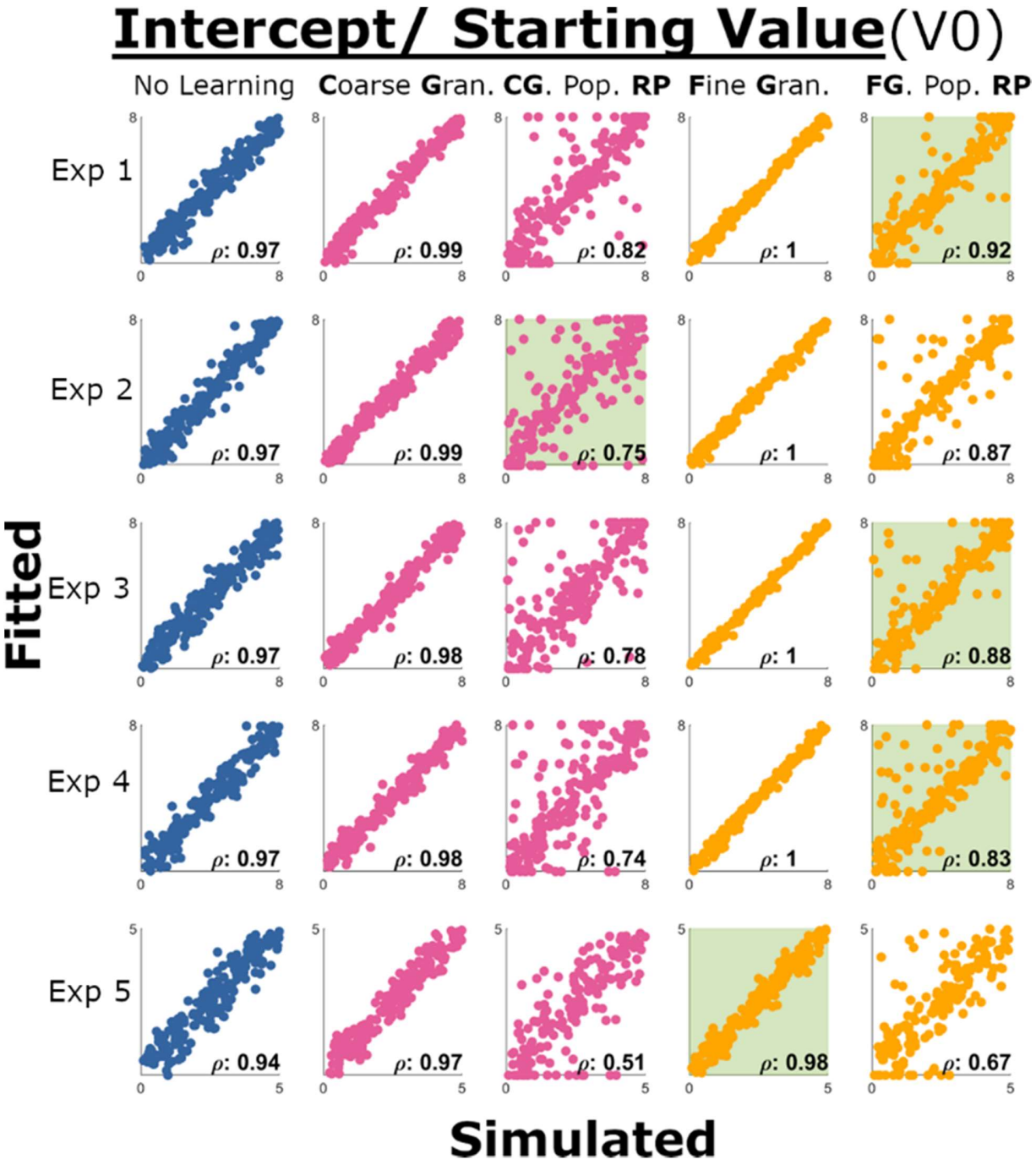


Figure 3.11. Parameter recovery for V0 (starting value).

See Figure 9, for an explanation of how we conducted the parameter recovery. Parameter recovery for the intercept (Model 1) and the starting value (Models 2-5) is acceptable for all models. Every row is an experiment and each column is a different model, green boxes indicate this was the winning model for that experiment.

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.3.7 Additional Models

In addition to the computational models discussed above we performed some exploratory analysis on some extra models. First, to see whether participants learned differently from positive and negative feedback we ran models 2-5 again but now with separate learning rates for positive and negative feedback (Figure 12). These models were worse for all experiments except for experiment 4 (Figure 12D). In this experiment the Fine Granularity and Stereotypical RP model with positive and negative learning rates performed best. The fitted parameters for this model indicated that participants learned more from positive feedback.

Finally, we tested all models again but with their self-ratings as RPs instead of the standard population average (Figure 13). In none of the experiments were these models better than the standard models with population based (i.e., average) RPs. Indicating that it was better to use the population based RP.

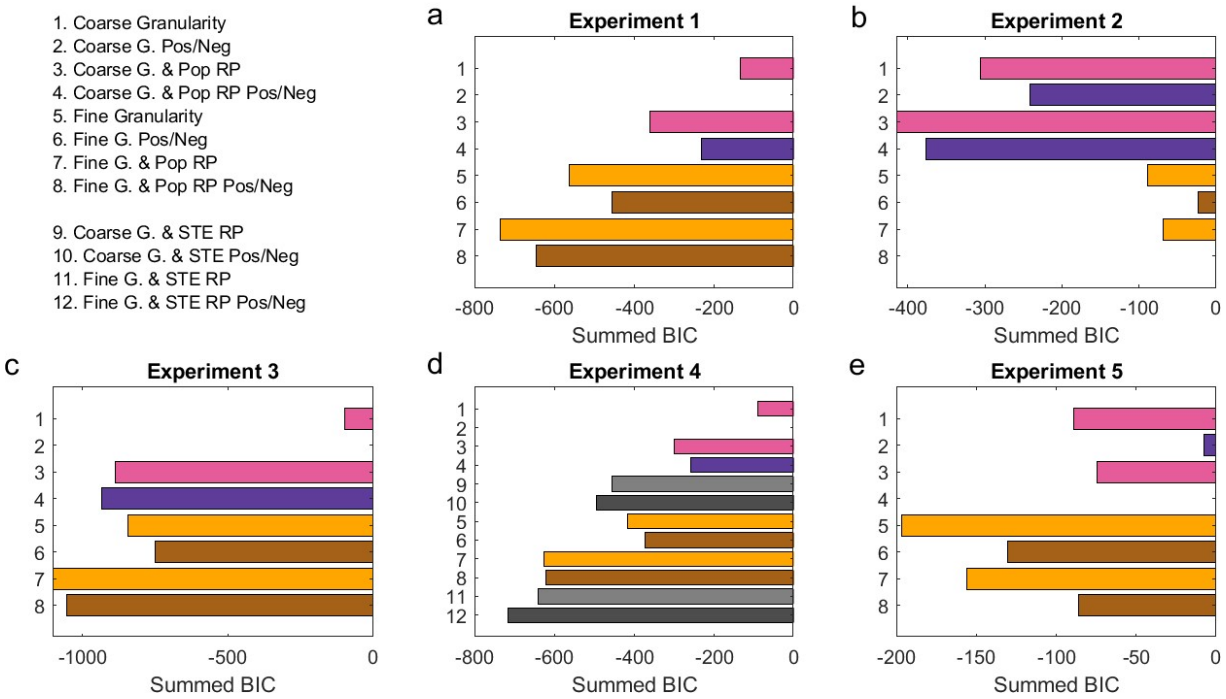


Figure 3.12. Positive/ Negative models (P/N) compared with the standard models.

Positive/ Negative models (Pos/Neg) have two parameters for the learning rate (α) by splitting it up for positive and negative PEs. This allowed us to check whether participants learned differently from positive and negative PEs. Each plot contains the regular models in standard colors. Below each regular model are those that make use of the positive/

negative learning rates (shown in darker colors, i.e., purple for the coarse models and brown for the fine models). For Experiment 4 we also included the stereotype models and their respective stereotype Pos/Neg models (these are depicted in light gray for the standard stereotype models and dark gray for the stereotype models that also use the Pos/Neg learning rates). CG = coarse granularity, FG = fine granularity, RP = reference point, PEs = prediction errors, STE = Stereotypical

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

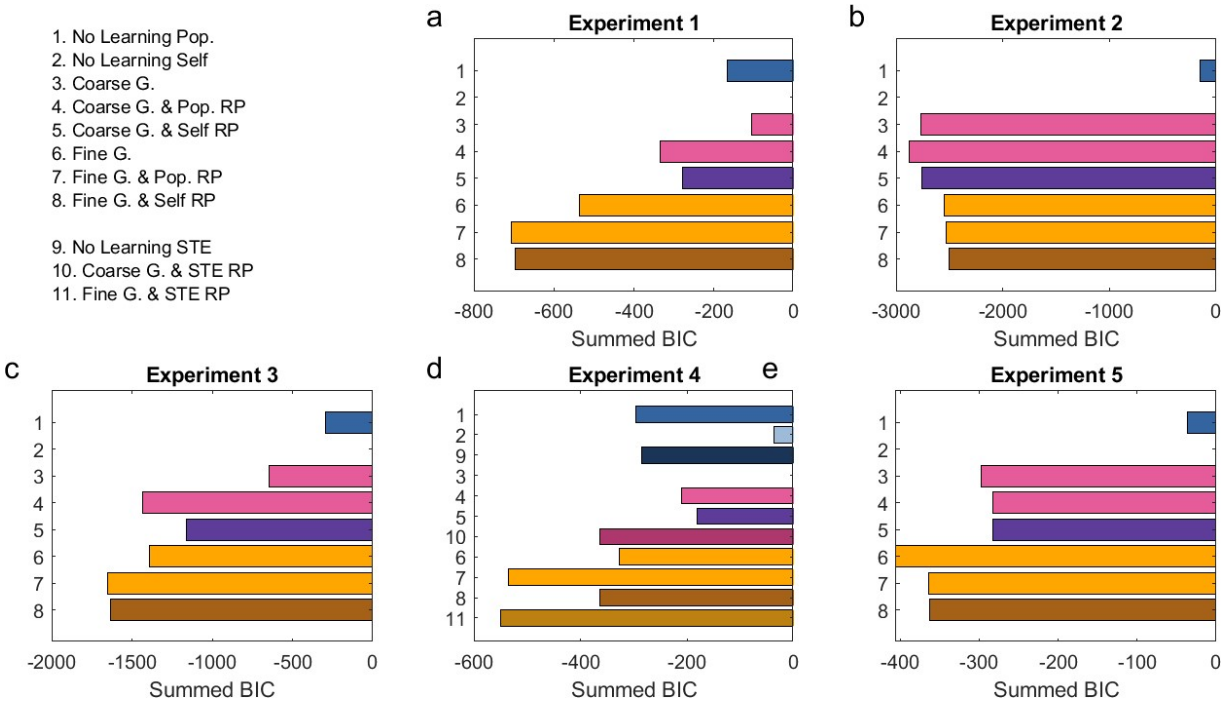


Figure 3.13. Self-ratings models (Self) compared with the standard models.
Model comparison of all models with additional models that use the participants' self-ratings as reference point (RP) instead of the population RP. The regular models are displayed in their original colors, directly below these are the models that use the Self RPs (in colors: light blue [No Learning], purple [Coarse], and brown [Fine]). For experiment 4 the same order holds but the STE models are added below the Self RP models in their original darker colors (see figure 6). Models that use self-ratings as their RP are always a little worse than those that use the population average RP. CG = coarse granularity, FG = fine granularity, RP = reference point, PEs = prediction errors, STE = Stereotypical

Figure from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

3.4 Discussion

In this chapter we examined how humans learn about others personality traits. We hypothesized through various computational models that humans would flexibly adapt their strategies based on the available information. We focus on two knowledge structures: granularity and reference point. Granularity indicates the level of detail with which one represents and updates the person they are learning about. Coarse grained means you only pay attention to the Big-5 factors and for fine grained, one pays attention to each trait individually as well as the relationships between items. In experiments 1-3 we found that participants switched between these two granularities based on the information at hand in the task. Evidence for the use of the reference point came from experiment four where participants used a stereotypical reference point instead of the regular student average which would have been the better option according to our simulations. Finally, in the last experiment we extended our models to also run on the learning data of different items. These modeling results were supported by standard statistical analyses and model simulations. Where the simulations focused both on the data as well as determining model robustness and distinguishability (Wilson & Collins, 2019).

The importance of this study is twofold. First, we discovered more details of how humans learn about others' personalities. That is, they seem to use detailed knowledge structures during learning. What knowledge structure gets used is dependent on the situation and the information available. Furthermore, the pervasiveness of stereotypes was exemplified in experiment four. Second, we introduce a novel modeling paradigm that can, in our opinion, be used for understanding more than just personality learning. We first alluded to this broader scope with experiment five that used a different set of stimuli but on which our models still fitted fine. Furthermore, a similar model space has already been used successfully on a preference learning task (Rosenblau et al., 2021). The reference points could be any preconceived information that one might have about the subject one is learning about. Similarly, granularity, which in its most detailed form is just a correlation matrix, can be calculated on almost any imaginable data. Of which data of similar form has been found in the human cortex (Peer et al., 2021; Stolier et al., 2020; Tamir et al., 2016). Furthermore, taking into account the brain's tendency to use the same solution for many problems it is not a far cry to assume it will use the same types of structures for other learning problems.

This does not leave this study without some limitations. First, the total amount of trait items per profile was low. In experiment five we only used 50 trait items (10 per factor). Adding more items would most likely increase the accuracy of the drawn conclusions. In our current task we use integer numbers to show how much of a personality trait a person possesses. Even though using numbers for feedback is used in performance evaluations in a real life setting one would more likely have verbal feedback and most likely fractional numbers. Future research should thus look at how these numbers are translated from numerical to verbal and deal with verbal feedback directly. Moreover, often, learning about others is based on actions from ourselves (e.g., approaching someone who seems kind) and others (e.g., judging someone who treats another unkind). Adding actions to the task would increase the external validity.

Future work should focus on the reference point and granularity. First, in this study we only use the two extremes of granularity (fine and coarse) the use of any other state on the continuum between these two should be explored. That is, how is the granularity learned in the first place and does it change over time? Similarly with the reference point it would be interesting to see how it is learned initially and how it changes over time. Finally, testing the reference point on different groups (i.e., different stereotypes) could result in valuable findings.

4. Study 2 – Patterns of activation during personality representation

4.1 Introduction

Humans can quickly learn by abstracting and generalizing information based on previously acquired knowledge structures. Such knowledge structures can be mathematically specified and understood as an aggregate of learned information. Previous research has found evidence for knowledge structures and their use during learning across a wide range of domains: In behavioral studies, knowledge structures have been shown to guide natural object perception (Hebart et al., 2020), language processing (Lenci, 2018) as well as learning abstract non-social (Wu et al., 2020) and social information . Neuroimaging studies have provided evidence for the representation of knowledge in neural patterns of activity across multiple domains such as language processing (Carota et al., 2017), learning abstract information (Garvert et al., 2023; Kahnt & Tobler, 2016), and in emotion processing (Skerry & Saxe, 2015). Here, we test whether social knowledge structures, i.e., aggregated knowledge of the relationships between personality traits and coarser personality dimensions, are neurally represented.

Personality traits capture a person's typical behavioral tendencies across a wide range of situations, thereby providing relatively stable generalizations of a person's behavior over time (Ashton, 2018). Knowing about a person's traits allows us to make accurate predictions about that person's future behavior (Epstein, 1979). Abstracting from personality traits to the broader multidimensional personality structure has played a large role in personality research. The number of dimensions needed to adequately describe personality structures has been a question of much debate in personality psychology. Some research findings show that two dimensions suffice to describe human personality (Fiske et al., 2002; Gray et al., 2007; Oosterhof & Todorov, 2008), while the influential five dimensional personality model, posits that five dimensions are needed (Goldberg, 1990; R. R. McCrae & Costa, 1987). This five-dimensional model, coined Big Five, encompasses the factors Agreeableness, Extraversion, Conscientiousness, Openness, and Neuroticism. Items within each factor are more closely related to each other than items between factors. That is, a person who is generally more agreeable will score higher on items of this factor

(e.g., helpful, generous, and sympathetic), thereby showing less variability in scores of items within that factor than on items within other factors.

While personality dimensions, lower-dimensional reductions of the personality trait space, have been shown to predict behavioral patterns, finer-grained trait structures afford more detailed representations of other persons (Frolichs et al., 2022). Such fine-grained structures can be formed by the relationships or similarities between single trait items. For example, learning about someone's helpfulness gives you more information about their generosity than their politeness. These fine-grained updates about another person may allow us to make more accurate predictions about specific characteristics and behaviors. We have previously investigated coarse- and fine-grained generalizations during personality learning (Frolichs et al., 2022). Akin to other research, which has shown that knowledge structures adjust based on their specific use (Thornton et al., 2023; Thornton, Weaverdyck, & Tamir, 2019) or co-occurrence (Bonner & Epstein, 2021), we found that depending on the task demands, participants would employ coarse- or fine-grained representations of personality. In line with previous research suggesting low-dimensional representations during social learning (King-Casas et al., 2005; Tamir & Thornton, 2018), we found evidence for the use of coarse granular representations. Interestingly, we also found evidence for the use of fine-grained representations when the task demanded for it. Using fine granular representations during personality learning allowed participants to use the similarities between individual trait words to leverage learning.

To summarize, representing social knowledge at different levels of granularity allows for flexible generalization during personality learning. We can solely rely on coarse-grained dimensions or personality factors to learn about overall tendencies of a person (i.e. coarse-grained learning) or take individual items and their fine-grained relationships into account for a more detailed representation and updating about the person in question.

A whole suit of analyses that fall under multivariate pattern analysis (MVPA), allow us to question the nature of these representations in cortical activity patterns. Similar to behavioral studies, neural evidence for social learning has focused on low-dimensional learning, such as low-dimensional personality characteristics (Delgado et al., 2005; Fareri et al., 2015; Jones et al., 2011; Tamir et al., 2016; Thornton & Mitchell, 2017; Zaki et al., 2016), or one or two dimensions of personality traits (Hartley & Somerville, 2015; Hassabis et al., 2014). Social decision making tasks recruit a dedicated network of cortical areas (Frith, 2007), which include the medial prefrontal cortex (mPFC), medial parietal cortex, temporo-parietal junction (TPJ) and the superior temporal

sulcus (STS) (Mitchell, 2008; Van Overwalle & Baetens, 2009). An area of specific interest to us is the mPFC, because it has been extensively implicated in computing decision variables in various task contexts (Bang & Fleming, 2018; Klein-Flügge et al., 2022), prediction errors (Gläscher et al., 2010; Joiner et al., 2017b), and accessing conceptual knowledge during decision making (Ghosh & Gilboa, 2014; Kumaran et al., 2009; Theves et al., 2021). Furthermore, a gradient within the mPFC from ventral to dorsal differentiates between self-versus other-referential processing. The ventromedial prefrontal cortex (vmPFC) has been more strongly implicated in self-referential processing (Amodio & Frith, 2006; Sul et al., 2015; Wagner et al., 2012) and the dorsomedial prefrontal cortex (dmPFC) has been more implicated in other-referential processing (Saxe, 2006; Sul et al., 2015). This self-versus other distinction, however, may in part be due to taking on a more active or passive decision making role during self and other-referential judgements.

Here, we aimed to investigate whether mPFC activity encodes finer grained as supposed to only coarse social knowledge by using a multivariate representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008). This technique allows us to compare representations from behavioral data with the representations from fMRI data. After its introduction, RSA has been widely applied in diverse topics and has shown its usability in deciphering neural activity patterns during social tasks (Peer et al., 2021; Popal et al., 2019; Riberto et al., 2022; Stolier et al., 2020). However, as mentioned above, so far, research on personality learning has only focused on low-dimensional representations of personality such as two of the Big-5 factors (Hassabis et al., 2014), or three general dimensions (Tamir et al., 2016). The finest representations so far found evidence for the Big-5 factors along with other lower dimensional representations (Thornton & Mitchell, 2017, 2018). However, the Big-5 model was the coarsest model we found evidence for in our previous study (Frolichs et al., 2022). We therefore hypothesized that, along with coarse, more fine-grained knowledge structures to be represented in the cortical activity patterns. Specifically, we expected to find these patterns in the mPFC.

To investigate whether more fine-grained patterns exist in cortical activity patterns we analyzed the data from two fMRI studies. Study 1 was a reanalysis of a previous study (Korn et al., 2012) that has been previously analyzed with a univariate fMRI approach only. The study consisted of a naturalistic social decision making task that involved real social interactions before fMRI scanning. Consequently, participants were immersed in a rich social context prior to making decisions about themselves or another person. The experimental design was conducive to RSA but the study was not optimized for this type of analysis. In study 2, we investigated social decision

making in a new sample of participants. This time the fMRI paradigm was optimized for RSA analysis (with respect to stimulus timing and repetition, and controlled for motor cortex activity) but consequently it was a less naturalistic task.

In both studies we asked participants to rate themselves and others on a number of trait words (e.g., generous & diligent). We extracted the neural patterns of activity during these task ratings using searchlight (SL) and region of interest (ROI) RSA. Furthermore, for study 2, we also got participants' explicit similarity ratings between items so we could create a model out of these idiosyncratic ratings. We hypothesized that the mPFC region contained both coarse and fine knowledge representations during personality learning given that we have shown that fine-grained trait similarity representations scale the integration of prediction errors on a trait learning task (Frolichs et al., 2022).

To test these hypotheses, we specified four models that varied the level of granularity in knowledge representations (from coarse to fine grained). As hypothesized, we found coarse and fine-grained social knowledge representation during social decision making in the dorsal medial prefrontal cortex (dmPFC). In study 1, we found more robust evidence using the searchlight whole brain analysis while in study 2, we found mixed evidence - the searchlight analysis revealed no significant clusters of interest, however, the a region of interest analysis revealed significant coarse and fine-grained knowledge representations in the dmPFC. This paper examines the validity of our results across studies and discusses potential reasons for the observed differences between the two studies.

4.2 Methods

4.2.1 Study 1 - Personality learning after previous social interactions

Data from this study were collected for a previous study (Korn et al., 2012). This study examined how humans process social feedback about their own and others' character traits. Importantly, this experiment was conducted in a real-life social setting. That is, the day before the scanner task, the participant along with four others played the board game monopoly to get to know each other's personality. Participants were aware during the game that they and the others would be rating each other's personality traits afterwards. Overall, results revealed a positivity bias in social feedback processing, more updating from positive as supposed to negative information.

In the original publication, data from this study was only analyzed with univariate analyses. However, the collected neural data lends itself to be reanalyzed to answer our novel multivariate questions using RSA. This analysis can reveal knowledge representations encoded in neural activity during this social decision making task. This study's only focus is the novel RSA analyses.

4.2.1.1 Participants

In total, 30 right-handed subjects participated. Three participants had to be excluded (one did not tolerate the scanner environment, another showed excessive head movement (>8 mm), and data from another subject could not be used due to technical problems) leaving 27 subjects for analyses (14 female, mean age = 24.3 years, SD = 2.46). All subjects gave written informed consent.

4.2.1.2 Scanner Task

Participants got to know and interact with four people the day before completing the social decision-making task in the scanner. In the scanner task, participants were asked to rate the character traits of one of the other four participants that they met on the previous day. On every trial, participants first saw a cue indicating whether they had to rate themselves or the other person (1 second). After this, they saw one of 80 trait adjectives and they had to imagine how much this trait applied to themselves or the other person (4 seconds). Afterwards, participants were prompted to rate on an 8 point Likert scale how much this trait applied to themselves or to the other person (6 seconds). After submitting their rating, participants received computer-generated feedback on their own traits or the trait ratings of the other person in question (Figure 4.1A).

4.2.1.3 MRI data acquisition

MRI data were acquired on a 3 T scanner (Trio; Siemens) using a 12-channel head coil. Functional images were acquired with a gradient echo T2*-weighted echo-planar sequence (TR = 2000 ms, TE = 30 ms, flip angle = 70, 64 × 64 matrix, field of view = 192 mm, voxel size = 3 × 3 × 3 mm^3). A total of 37 axial slices (3 mm thick, no gap) were sampled for whole-brain coverage. Imaging data were acquired in four separate 349 volume runs of 11 min 38 s each. The first five volumes of each run were discarded to allow for T1 equilibration. A high-resolution T1-weighted anatomical scan of the whole brain was acquired (256 × 256 matrix, voxel size = 1 × 1 × 1 mm^3).

4.2.1.4 Pre-processing

Image analysis was performed using SPM8 (www.fil.ion.ucl.ac.uk/spm). Echoplanar imaging images were realigned, unwarped, and coregistered to the respective participant's T1 scan; normalized to a standard T1 template based on the Montreal Neurological Institute (MNI) reference brain; resampled to 3 mm isotropic voxels.

4.2.2 Study 2 – Personality inferences in the absence of social interactions

Data for this new study was collected after a preliminary analysis for study 1 with an RSA SL analysis. Given that the scanner task in Study 1 was not specifically designed and thus not optimized for RSA analysis, in Study 2, we sought to replicate our results in a more controlled task that was specifically designed for RSA. To increase the robustness of the task for multivariate analyses, we added more scanning sessions (from one session in study 1 to five sessions in study 2). To keep the total scanning time within a reasonable timeframe, we removed the social feedback phase. This meant that participants rated themselves or a fictitious other on trait adjectives and no longer received feedback. To reduce unwanted activity in the motor cortex, we also reduced the answer options from an 8-point Likert scale to binary options (0: does not apply to me/ the other person, 1: applies to me/ the other person). Finally, to reduce the overall time of the experiment, participants only gave ratings in the scanner about a fictitious other that corresponds to the participant's notion of an average person (i.e., average person). Importantly, there was no social interaction session before or during the experiment. Importantly, we asked participants to rate the

similarity of trait words after the scanner task. This provided us with similarity ratings between all the trait words, which were used for the RSA model 4.

4.2.2.1 Participants

We recruited 32 participants through online advertisements. We aimed for a sample size similar to study 1 (n=30) (Korn et al., 2012). Participants had to meet the following inclusion criteria: 1) age between 18 and 40 years, 2) German native speakers, 3) normal or corrected to normal vision, and 4) no history of neurological or psychiatric disorders. The study was conducted in accord with the Declaration of Helsinki and approved by the local research ethics committee (Ethik-Kommission der Ärtzekammer Hamburg, Number: PV5746). All participants gave written informed consent using a form approved by the ethics committee and were compensated on an hourly basis.

4.2.2.2 Behavioral data acquisition

Scanner Task

The scanner task consisted of five identical runs that were all completed in one scanning session. Participants were instructed before placement in the MR scanner. A single run consisted of 100 trials and during every trial one trait word was presented. In total, there were 40 unique trait words. Participants were asked to rate, with "yes" or "no", if the trait word described either themselves (i.e., you-trials, n=40) or an average person (i.e., other-trials, n=40), the remaining 20 trials were null trials that only showed the fixation cross for a duration of 5 seconds. Participants were shown the same 40 trait words for themselves and the average person. Both the trait words and the trial type (i.e., you- and other-trials) were randomized within each run. Each trial lasted exactly 5 seconds.

In more detail, every trial started with a 250ms fixation cross, after which the trait word (e.g., diligent) was shown for 2750ms (imagination phase). Then two answer options (Yes, No) were presented underneath the word on the same height (decision phase). The position of these words was counterbalanced on the screen (left and right) and subsequently as an answer option on the button box. During this decision phase participants were required to respond within a 1500ms time interval otherwise the warning ("Too slow!") appeared on the screen and no answer would be recorded. If participants made a response within this decision phase, they were able to see their choice until the full 1500ms duration had elapsed. To make sure participants saw their answer or

the warning from the decision phase, the responses remained highlighted on the screen for another 250ms after the end of the decision period. A final fixation cross of 250ms duration ended each trial (Figure 4.1B).

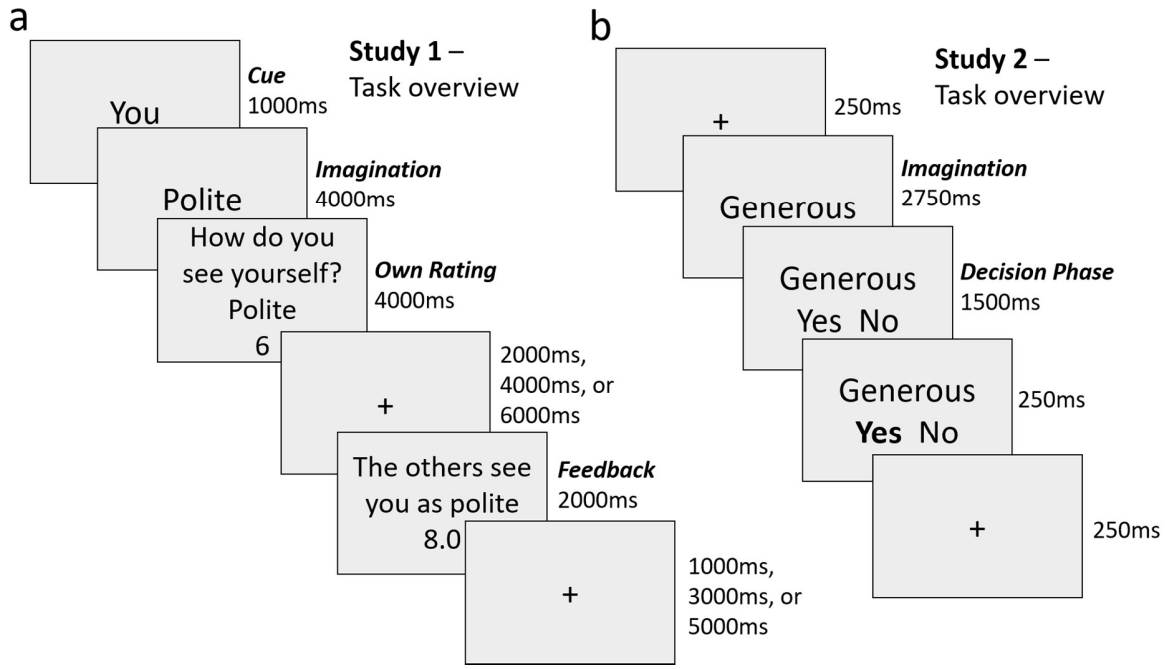


Figure 4.1. Task overview for both studies.

Both studies had similar tasks where they were asked to rate their own and other's personalities. For study 1, they rated a person they interacted with before and for study 2, an average person they had never met and would never meet. For study 1, participants rated on a scale of 1-7 (1 – does not apply at all, 7 – applies completely), whereas for study 2 ratings were binary ("no" – does not apply, "yes" – does apply) during the scanner task but more detailed (1-7 scale) afterwards. In study 1, participants had slightly more time for the imagination and rating phase and ended each trial with a feedback phase. The imagination and rating phases in study 2 were a bit shorter and there was no feedback.

Post Scanner tasks

After scanning, participants performed several rating tasks of the previously seen trait words (N=40) and previously unseen trait words (N=85). All ratings were made on a scale of 1-7 (i.e., 1: not important/ does not apply, 7: very important/ applies very much). First, they rated how much 125 traits words applied to themselves as well as to an average stranger that they have never met (see the Supplementary Table 1 for a complete list).

For the 40 trait words that were used in the scanner task (Supplementary Materials table 2), participants additionally performed several other ratings. For these trait words they were asked 1) to rate how positive they found each trait word, 2) how important it was for themselves or someone

else to possess this trait, and 3) how easy it was to rate themselves or an average person on each trait.

Finally, we asked participants to explicitly rate all possible trait pairs (i.e., 780 unique pairs) based on their similarity on a scale of 1-7 (i.e., 1: not similar at all, 7: completely similar). For example, helpful and polite are similar qualities and should be rated as being similar). There was no time limit on these individual ratings and participants were free to take breaks as they needed them.

4.2.2.3 MRI data acquisition

We acquired T2*-weighted functional images on a 3 Tesla Siemens Magnetom Prisma with the following parameters: repetition time (TR): 1313ms, echo time (TE): 30ms, FA: 50°, FoV: 224 mm, number of volumes: 384, number of slices: 42, slice thickness: 2 mm with 25% (0.5 mm) gap for a total acquisition time of: 504.2 seconds. To correct for distortions we acquired a field map with the following parameters: TR: 518ms, TE1: 6.12ms, TE2: 8.58ms, FA: 40°, FoV: 224 mm, slice thickness: 2 mm.

We acquired T1-weighted anatomical images using magnetization-prepared rapid gradient echo sequence (MPRAGE) with the following parameters: TE: 2.98ms, TR: 2300ms, FA: 9°, FoV: 256 mm, slice thickness: 1 mm.

4.2.2.4 Pre-processing

fMRI data were pre-processed using SPM12 software (www.fil.ion.ucl.ac.uk/spm/software/spm12/) in MATLAB (version: r2020b). We performed slice time correction, corrected for signal distortions using the field map, and realigned functional scans to the first image using a six-parameter rigid body transformation to correct for motion. The structural image was co-registered to the functional image, and was segmented and normalized to MNI space. After this the realigned functional images were normalized using the transformation parameters from the structural image normalization, and smoothed using an 6 mm full-width at half maximum Gaussian kernel.

4.2.3 Neuroimaging Analysis – Both Studies

All neuroimaging analyses were conducted using SPM12 and custom code written in MATLAB (version r2020b). Within MATLAB we further used the Marsbar toolbox (Brett et al., 2002) for creating regions of interest and the RSA toolbox (Nili et al., 2014) for some of their functions.

All neuroimaging figures were made using MRIcroGL (version 20-July-2022) with the following procedure. Resulting activations were first displayed with SPM using the following cut-off: surviving a threshold of $p < 0.05$ after clusterwise familywise error correction for multiple comparisons over the entire brain at a cluster-defining threshold of $p < 0.0001$ (uncorrected). The resulting T-maps were then imported in MRIcroGL and brightness and darkness thresholds of the t-values were adjusted to fit those of SPM as close as possible. Finally, the coordinates of the slices were taken from SPM's global or local peak voxel within the cluster of focus.

4.2.3.1 Univariate analyses

We performed two univariate analyses for Study 2. The first GLM, focused on the broad differences in activity between self and other. That is, for each session, the self-imagination-phase and other-imagination-phase trials were modeled as separate conditions beginning with the trial word onset for the duration of the trial (2750ms). Six motion regressors of no interest were added for each session as well. Both contrasts for self > other and other > self were tested.

These analyses intended to confirm that the mPFC was active during this social learning task as expected from the social learning literature (Frith, 2007). The second analysis functioned as a sanity check for the later RSA searchlight analysis. From the literature, we generally expected higher activity during thinking about self, compared with activity during thinking about others. As a sanity check, we compared activity in the visual cortex from the univariate and an RSA SL analysis.

The second GLM functioned as the first part of a sanity check of our later analyses. We expected a robust effect in the visual cortex based on the length of the trait words i.e., more activity for longer trait words. To do so, we modeled this GLM with a parametric modulator for the length of each trait word. That is, for each trait word separately we modeled a parametric modulator on the number of letters in that word. Six further regressors of no interest were added to each session to model participant motion.

These same word length differences can also be captured in a model to use with an RSA SL analysis (explained in more detail below). Results from the second GLM can then be compared with the results from the RSA SL analysis. Because these two analyses are distinct in their setup but focus on the same patterns of activity we can compare their results. We take the univariate analysis as our ground truth and compare if these two analyses reveal similar areas of activity. When they do this supports that our RSA SL analysis functioned as intended.

4.2.3.2 Representational Similarity Analysis

To test to what extent social structures are represented in the brain we conducted representational similarity analysis (RSA) using both an ROI and a searchlight analysis. Since the two studies had a different set of stimuli, calculation of the parameter estimates (i.e., beta values) will be described separately.

Study 1

The beta values for study 1 were calculated with a standard GLM. We modeled the presentation of each trait word as a regressor of interest for 160 total regressors (80 self- and 80 other-imagination-phase, 4000 ms), together with 6 regressors of no interest to correct for motion. This resulted in a single beta image for each trait word. On all these beta images we calculated a contrast, resulting in as many contrast images as original betas. These contrast maps were used for further RSA analysis.

Study 2

To calculate the beta values for study 2 from the GLM we modeled the presentation of each trait word as a regressor of interest (80 regressors, 40 self- and 40 other-imagination-phase, 2750 ms) for each session. Six further regressors of no interest were added to correct for motion. This meant that we obtained a single beta image for each trait word in each session separately for the self- and other-imagination-phases. To reduce noise we averaged these beta images over the five sessions across trait words to create 80 average beta images (40 for self and 40 for other). Each analysis described below was run separately for the self and other beta images.

For both studies, we first performed a searchlight analysis. Depending on the results of this analysis we performed a region of interest analysis thereafter.

RSA - Searchlight

For this analysis we created a spherical volume (5 voxel radius (similar in size to other social learning studies which used a 3 voxel radius: (Peer et al., 2021; Riberto et al., 2022), for a total of 515 total voxels) that we moved throughout each of the participants' brains (i.e., beta maps) centering on each voxel. From each of these positions we extracted the patterns of activity (i.e., the beta values) from all the voxels in the spherical volume. From the activity within these voxels

across all trait words we calculated a data RDM using the Pearson Correlation coefficient. The resulting matrix was subsequently subtracted from 1 to create a data RDM. Next, we calculated to what extent the model RDMs correlated with the data RDM. This correlation between the data RDM and each of the models was calculated with Kendall's Tau alpha. This correlation coefficient value was placed on the center voxel after which the SL moved on to the next center voxel and repeated the previous steps until there was a whole brain correlation map.

The correlation coefficient brain maps from all participants were then taken into a group comparison (i.e., second level analysis in SPM). All reported activations survived a threshold of $p < 0.05$ after clusterwise familywise error correction for multiple comparisons over the entire brain at a cluster-defining threshold of $p < 0.001$, uncorrected.

To test the searchlight analysis procedure we created a model RDM that represented the letter length of each word that was presented. We visually compared the results from this searchlight analysis with the results from a univariate analysis that had word length as a parametric modulator.

RSA - Region of Interest

In the ROI analysis we tested the similarity in activity patterns between two specific brain regions (see section Regions of interest) and three dissimilarity models (see section Dissimilarity Models) during self- and other-imagination-phases. All analyses were conducted separately for these self and other-imagination-phases.

From each of the beta images (80 for study 1 and 40 for study 2) we extracted the voxels within the ROI and calculated the pattern dissimilarity using the Pearson correlation coefficient between the neural activity patterns of all trait words to get a correlation matrix that contained all correlations between activity patterns for the trait words. This correlation matrix was subtracted from 1, which resulted in two matrices (i.e., the data RDMs for self and other-ratings). These data RDMs were compared with each of the three model RDMs for each participant using a rank correlation (i.e., Kendall's Tau alpha, from the RSA toolbox (Nili et al., 2014)). To determine if a model significantly represented the neural activity pattern we calculated a one-sided t-test against zero for these individual rank correlations. All alpha values were Bonferroni corrected for the number of models within each analysis.

Regions of interest

To test patterns of activity in specific regions, two regions of interest (ROIs) were created. The first was determined from the functional activity of the first univariate GLM analysis from study 2 (contrast self > other). This GLM revealed a large cluster in the (v)mPFC (see Figure 5, blue cluster). The second cluster was taken from the SL analysis of study 1. This searchlight analysis revealed a cluster of activity in the dmPFC (see Figure 5, yellow cluster). Both these functional clusters were extracted as a mask using the Marsbar software package (Brett et al., 2002; marsbar.sourceforge.net/) and any subsequent ROI analyses were performed on these two ROIs.

4.2.3.3 Dissimilarity Models

Because we are interested in exactly how detailed the personality representations are in the cortex during personality learning, we created four distinct models that all represent personality but on differing granularities. We hypothesized that the models would explain the activity patterns in order from coarse- to fine-granularity. Note that the coarse granularity structure is also present in the fine granularity models. We hypothesized that coarse models will significantly explain neural activity patterns and we predicted to find neural representations of fine-granularity knowledge structures, particularly in the MPFC.

In brief, Model 1, was the least complex Big-5 model, which assumes that patterns of brain activity reflect the coarse grained Big-5 factor structure as previously shown (Thornton & Mitchell, 2018). Model 2 tested fine-grained similarity representations. Correlations between self-ratings of individual items were expected to scale with patterns of neural activity. This model is based on an independent group of participants. In a previous study, we found that this fine-grained representation was used in personality learning (Frolichs et al., 2022). Model 3, tested fine grained similarity representations using explicit similarity ratings, namely participants' idiosyncratic similarity ratings. This model states that activity patterns reflect participants' own explicit conception of how similar items are. We classify this model as finer-grained than Model 2, because each participant has their own individual similarity model. For models 1 to 3 we hypothesized that the models would significantly explain the activity patterns in order from coarse to fine (i.e., if any model is significant, we expect the coarse model 1 to explain the data significantly first followed by models 2 and 3). We expected to see this because the coarse information structure from model 1 is also captured in the more complex models. We additionally tested a non-social visual perception model as a control test of our analysis pipeline Model 4, was a word length similarity

model, which modeled the absolute length differences of the stimulus words. We expected that perceptual differences between trait words, i.e. their length, would be represented in the primary visual cortex. We further confirm these results with a standard univariate analysis.

Model 1 - Big-5

The coarsest of the models, the **Big-5 model** (Figure 4.2B, E), is a conceptual model that indicates what trait items belong to the same factor. That is, it assumes patterns of activity are the same within each factor (i.e., if two traits belong to the same factor they will have the same activity patterns). This model uses binary options where "1" indicates a trait item is part of a factor and "0" indicates it is not. To create the RDM we subtracted the model from 1, this meant that items within a factor had a dissimilarity rating of "0" (i.e., not dissimilar) whereas items between factors had a dissimilarity of "1" (i.e., completely dissimilar). We created separate Big-5 models for study 1 (size 80x80) and study 2 (size 40x40). These items were taken from a large list of German trait adjectives (also used in previous studies (Korn et al., 2012; Korn, Sharot, et al., 2014)) and were hand-annotated to fit the Big-5 factors (as in our previous work (Frolichs et al., 2022)). Originally the items were not sorted for the Big-5 factors, Figure 4.2F shows the Big-5 model when items are sorted for the factors.

Model 2 - Similarity

We created two similarity models, both these models use self-ratings from a sample gathered for the first study (Korn et al., 2012). These models assume that patterns of activity in the cortex show a similar relation to the correlations between the trait word self-ratings.

We created two different models from these self-ratings so we could answer a secondary hypothesis. This hypothesis focusses on how the cortex deals with positive and negative trait words. Normally, negative items (e.g., cowardly and spiteful) tend to get rated lower than positive items (e.g., tidy and respectful) we wondered if this difference in ratings is reflected in the cortical activity patterns. That is, does the cortex code differently for positive and negative items or does it code in the same scale, disregarding whether the item is positive or negative. The first of these two models, **similarity model 1-different** (Figure 4.2), keeps this difference in ratings between the positive and negative items and thus hypothesizes that cortical activity patterns are different for positive and negative items. In contrast to this, the second model, **similarity model 2-same** (Figure 4.2), hypothesizes that positive and negative items elicit the same cortical activity patterns that does not

take the item valence into account. To achieve this, we reverse-coded all the negative self-ratings (e.g., if someone rated themselves to score a "1" on spiteful, the lowest score, this got transformed to an "7", the highest score). This reverse-coding of negative items was reflected in the correlations, in that positive and negative items showed more relatedness.

For both the reverse-coded and the non-reverse-coded self-ratings the procedure was the same. First, these self-ratings were Pearson correlated to create a similarity matrix that showed the correlation between each trait word pair. These matrices were then transformed into representational dissimilarity matrices by subtracting them from 1. This resulted in distances from "0" (i.e., perfect correlation) to "2" (i.e., anticorrelated). Because both studies had a different number of stimuli we created separate similarity models for study 1 (80x80 similarity matrix) and study 2 (40x40 similarity matrix). A comparable similarity matrix based on the same sample has previously been calculated and used in another study from our lab (Frolichs et al., 2022).

Model 3 - Explicit Similarity Ratings

The e**xplicit similarity ratings model** (Figure 4.2C), was only used for Study 2 and was based on real idiosyncratic similarity ratings. This model assumes that relations between cortical patterns of activity are different for each individual and reflect their own specific and idiosyncratic similarity ratings. We asked participants to give their explicit similarity ratings [1-8] between all possible trait word pairs (i.e., 40 trait words for a total of 780 combinations). For example, participants were asked to rate, on a scale of "1" (completely dissimilar) to "7" (the same), how similar they found the two trait words "easy-going" and "cold-hearted". When using this model we tested each participants' own idiosyncratic and explicit ratings between the trait words instead of relying on an aggregate model such as the other two models. Thus, this model hypothesizes that each participant has distinct activity patterns that show similarities akin to their subjective explicit similarity ratings.

Model 4 - Word Length

The **word length model** was created to check our multivariate (RSA) analysis against a more standard univariate analysis for the second study. This model focused on the length of the stimulus words, this was done by taking the absolute difference between each of the stimulus words. That is, for each row in the matrix we took the absolute difference between all trait words and the trait word on the identity line. This resulted in a 40x40 matrix where the diagonal was all zeros. This

model hypothesized that longer stimulus words would elicit more pronounced activity than shorter stimulus words.

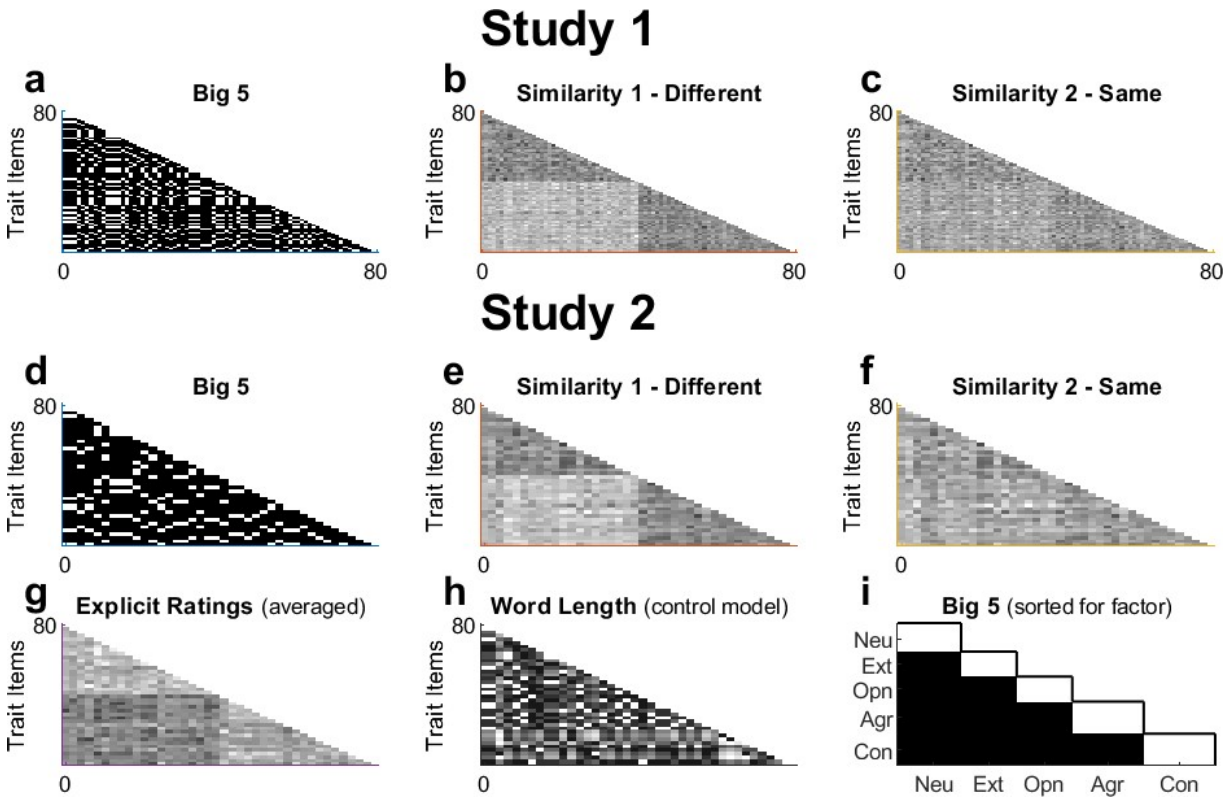


Figure 4.2. Model RDMs used in this experiment.

We use five different models in total for both ROI and searchlight analyses. Models colors on the axes correspond to the colors in the results for the ROI analysis (Figure 5). Study 1 used three models and study 2 used four. A final model (word length) was used to test the correctness of our searchlight procedure. The three models that are both used in studies 1 and 2 are conceptually the same but have more items for study 1 (80) than study 2 (40). **a- d**). The Big-5 model is a conceptual model (0 not in the factor, 1 in the factor). This model hypothesizes that activity patterns for items within the same factor are similar and different between factors. Because items are ordered according to positive and negative these factors are shuffled. **i**) Shows the Big-5 model sorted by factor. **b, e**). The similarity-1 Different, is the Pearson correlations between trait item self-ratings. These ratings have differences between positive and negative items which is reflected in the correlations. This model hypothesizes that activity patterns between items correspond to the correlations found in behavioral data, and that the patterns between positive and negative items are different. **c, f**). The similarity-2 Same, is the Pearson correlations between trait item self-ratings. However, the negative items from these ratings have first been converted to be in the same scale as the positive items. So that this model hypothesizes that activity patterns between items correspond to the correlations found in behavioral data, and that the patterns between positive and negative items are similar. **g**). The explicit ratings model, uses the explicit similarity ratings from the participant. In the analyses we used each participants' own ratings as the model. This model assumes that activity

patterns reflect those of the idiosyncratic ratings from each participant. **h**). The word length model uses the absolute differences in the trait word length. This model hypothesizes that the differences in word length are reflected in the cortex. We used this model to determine if our searchlight functioned as intended.

4.3 Results

4.3.1 Study 1 – Personality learning with social interactions

For study 1, we ran the searchlight analysis with three model RDMs i.e., a coarse model, the Big-5 factors (Model 1) and two fine models (similarities) with the individual correlations between trait words (Models 2a and 2b) (Figure 4.2).

4.3.1.1 RSA-Searchlight

We tested the above-mentioned models on two different phases within the scanner task, the self-imagination-phase, which is when participants were rating their own personality and the other-imagination-phase, which is when they were rating another person's personality. We hypothesized that we would mainly find regions in the mPFC for models 1-3 and the motor cortex.

Self-imagination-phase

Results during the self-imagination-phase showed a significant cluster for the Big-5 model in the motor cortex extending into the somatosensory cortex. This coarse model assumes activity patterns within factors are more similar than between factors. The second similarity model [similarity 2-same], revealed a cluster in the dmPFC (Figure 4.3A) and two larger clusters in both hemispheres of the motor cortex (Figure 4.3B) as hypothesized. This model presumes that the similarity of activity patterns between trait words reflects the correlations between trait words and that these patterns are similar for positive and negative trait words. For the first similarity model [similarity 1-different] we found clusters in the motor cortex and the medial temporal lobe. This model assumes that activity patterns between trait words reflect those between behavioral correlations and that the differences in activity between positive and negative items are larger.

A final searchlight analysis that focused on the length of the stimulus words [model 4 – word length] revealed two clusters in the bilateral visual cortex (Figure 4.3C). This model focuses on the visual difference on the screen caused by different stimulus word lengths and was thus hypothesized to show activity in the visual cortex.

Other-imagination-phase

We found less patterns of activity during the other-imagination-phase. For the Big-5 model, which assumes patterns of activity are similar for items within factors and different between factors, we did not find any significant clusters. The searchlight for the second similarity model [similarity 2-same] revealed two clusters, one in the bilateral motor cortex and the other in the cerebellum. For the first similarity model [similarity 1-different] there was a significant cluster in the premotor cortex (table 1). These similarity models assume cortical activity patterns reflect the correlations between items from behavioral ratings. Model 1-different assumes activity patterns are different between positive and negative items and model 2-same assumes these activity patterns are similar.

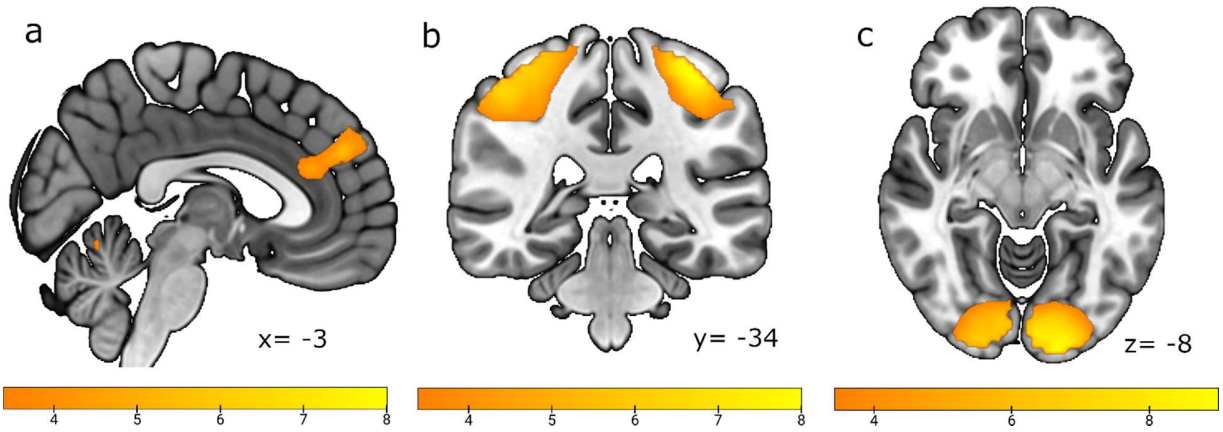


Figure 4.3. Results from the searchlight analysis for study 1 during self-ratings.

Results for the similarity model during the self-imagination-phase reveal **A**) a cluster in the dmPFC, where we expected to find activity and **B**) activity in the motor cortex. We tried to prevent this activity by adapting our paradigm in study 2. Instead of getting ratings on a scale of 1-8, which used four digits on each hand and thus evoked a lot of motor activity, we asked participants to rate on a binary scale ("Yes" - applies and "No" - does not apply) and counterbalancing the options so less motor activity would be present in the results. **C**) Results from the searchlight analysis for the length of the stimulus words. We expected significant clusters in the visual cortex.

	Side	Peak voxel MNI coordinates (mm)			Cluster size	p (cluster FWE corrected)	Peak z Score
		x	y	z			
motor cortex	L	-36	-34	64	779	<0.001	6.04
motor cortex	R	33	-19	67	1325	<0.001	5.4
mPFC	L/R	6	47	28	353	<0.001	5.33
motor cortex	L/R	-36	-13	61	4589	<0.001	5.7
medial temporal lobe	L/R	-21	-37	-23	851	<0.001	4.59
somatosensory	L	27	-31	55	273	<0.001	4.25
motor cortex	L/R	6	-31	43	3441	<0.001	5.09
Cerebellum	L	-18	-58	-26	218	<0.001	3.84
Premotor	L	-45	-37	58	823	<0.001	4.68

Table 4.1. Results from the searchlight analysis for study 1.

4.3.2 Study 2 – Personality learning with no social interactions

4.3.2.1 Neuroimaging Analysis

RSA

First, like in Study 1, we used a searchlight to find clusters in the brain that code in patterns of activity that encoded the model RDMs. We then performed ROI RSA analyses for the hypothesized MPFC region. These analyses tested the four different RDM models.

RSA – Searchlight

The regular searchlight analysis using all four standard models (i.e., the Big-5, both similarity models, and the explicit similarity ratings) did not yield any significant clusters. We next performed an analysis check to make sure our searchlight analysis functioned as intended.

Analysis check: Searchlight vs Univariate analysis

To make sure our searchlight analysis functioned in the way we intended we compared its results with a robust univariate analysis. That is, the effect of word length on activity in the cortex. So we ran two analyses, first a univariate analysis where we modeled a parametric modulator with the length of each stimulus trait word, from this analysis we expected strong activity in the visual cortex. Second, we performed a searchlight analysis with model 4 [word length] that modeled the absolute differences between word lengths, we also expected visual cortex activity with this model. If the clusters of activity between these two models overlap, we take this as evidence for successfully implementing the searchlight analysis. Both the univariate analysis (Figure 4.4, red shading) and the searchlight analysis (Figure 4.4, green shading), found a large cluster of activity in the visual cortex, these clusters showed significant overlap (Figure 4.4, yellow shading) confirming that our searchlight analysis worked as we had intended.

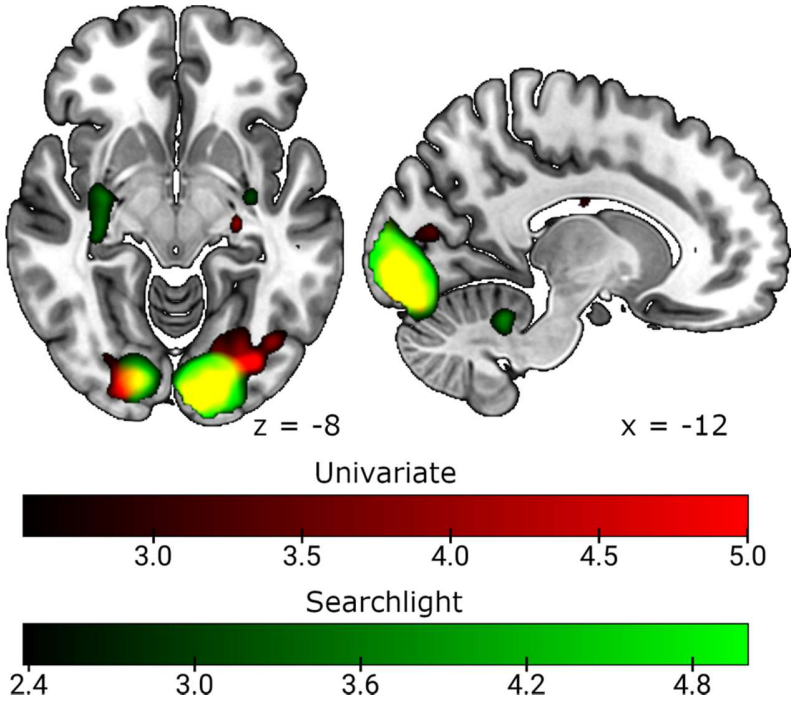


Figure 4.4, Univariate contrast and searchlight analysis for the length of the stimulus trait words.

Red shows the results from the univariate contrast, which had a parametric modulator for the length of each stimulus trait word. Green shows the results from the searchlight analysis with the word-length model [model 4]. Yellow shading shows the overlap between these two analyses. Because the two analyses show a high degree of overlap, we feel more confident that our searchlight analysis was performed correctly.

RSA - Region of Interest

For the ROI analyses we used two clusters that we based on previous analyses from this study. The first ROI was created from a univariate analysis from the second study (see Supplementary Analysis). In this analysis we ran a standard GLM on the trait word stimuli and contrasted the activity for the self-imagination-phase with the activity for the other-imagination-phase. This contrast returned a cluster in the mPFC (Figure 4.5, blue region). Because we already hypothesized the importance of this region, we used this as our first ROI. We then created a second ROI based on the searchlight results in study 1. The searchlight analysis for the model [similarity 2-same] during the self-imagination-phase resulted in a cluster located in the dorsal mPFC (Figure 4.5A and Figure 4.5, yellow region). Which was also hypothesized as a region of interest. We ran the four models (i.e., the Big-5, both Similarity models, and the explicit similarity ratings) on both these regions for both the self-imagination-phase (i.e., when participants were rating themselves) and the other-imagination-phase (i.e., when participants were rating a stranger). We expected that

the first ROI in the ventral prefrontal cortex would show more self-focused results and the dorsal ROI to show more other-focused activity patterns.

Self-imagination-phase

Testing four different models on the activity patterns during self-ratings of personality traits, we found that only the Big-5 model significantly explained the data for both ROIs. That is, the vmPFC (Figure 4.5, blue box top) and the dmPFC (Figure 4.5, yellow box top). The other models (i.e., similarity 1-different, similarity 2-same, and the explicit similarity ratings models) did not explain the activity patterns significantly well in both of the ROIs. That is, we only find evidence for the use of coarse knowledge structures during self-ratings of personality.

Other-imagination-phase

We tested the same four models during the personality ratings of strangers. Like during the self-imagination-phase we found the Big-5 model to significantly explain the patterns of activity. Furthermore, we found that both the similarity 1-different and the explicit similarity ratings models to also explain the patterns of activity significantly well (Figure 4.5, yellow and blue box bottom). Indicating that cortical activity patterns during other-ratings exhibit patterns of fine knowledge structures.

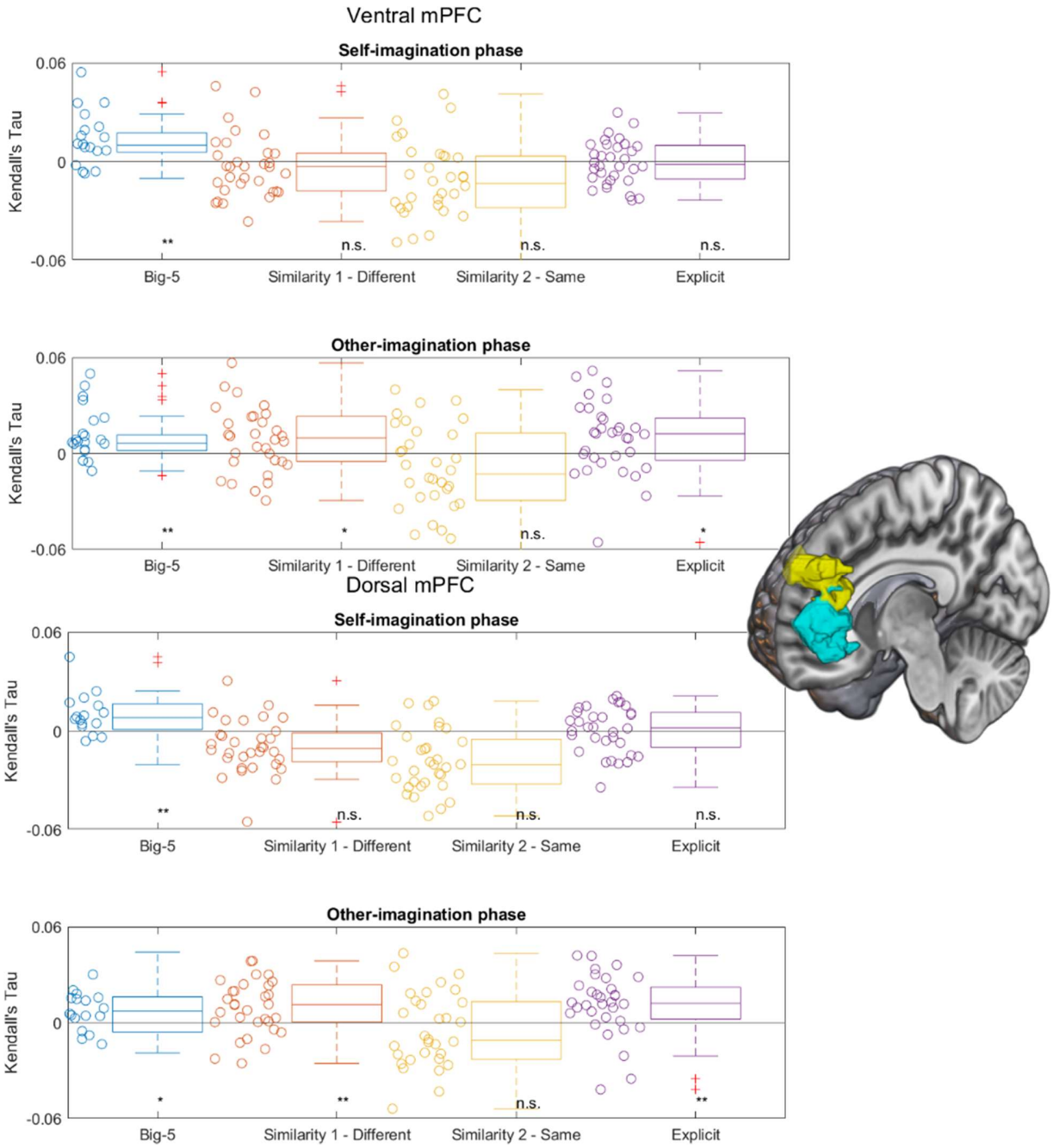


Figure 4.5. Results from the first ROI RSA analysis for four different models (Bonferroni corrected).

The main difference we find is between the self- and other-imagination-phases. That is for both ROIs we find that during self-imagination-phase only the coarse Big-5 model explains the patterns of activity significantly. The Big-5, Similarity 1 (Different), and explicit models explain patterns of activity well during the other-imagination-phase for the vmPFC. In the second ROI (dmPFC), we also find that the Big-5, Similarity 1 (Different), and explicit models explain the neural activity patterns well for the other- imagination-phases.

Individual data points are participants' parameter estimates which are summarized by boxplots (median (middle line), 25th, and 75th percentile (box), the whiskers extend to most extreme data points not considered outliers (1.5 times interquartile range), outliers are indicated with + signs)

4.4 Discussion

In this study, we set out to investigate social knowledge representations during social decision making with two studies and tasks. Study 1 contained naturalistic interactions followed by a social feedback task. Study 2 was a simplified and more controlled than Study 1. This task was specifically designed for RSA analysis, but also reduced participants' social context. Across studies, we expected to find patterns of neural activity that reflected social knowledge representations. This expectation was based on our previous behavioral studies that highlighted the use of social knowledge structures in social decision making tasks (Frolichs et al., 2022; Rosenblau et al., 2018). We tested both coarse and fine-granularity structures of personality traits. We hypothesize that both types of structures are encoded in the cortical activity. Evidence for simpler knowledge structures such as the Big-5, which hypothesizes that patterns of activity for trait words are similar for trait words within a factor and dissimilar between factors, has been found in previous research (Thornton & Mitchell, 2018) but, so far, no evidence has been found for more complex knowledge structures. We created four models that represent knowledge structures at different levels of granularity and used representational similarity analysis to find which of these models could explain the patterns of activity in neocortical areas and within the MPFC.

In the first study, we found strong evidence for both coarse and fine-grained social knowledge representation in the dmPFC during social decision making, specifically the self-imagination-phase. This model hypothesizes that patterns of activity are similar to the correlations between these items and that the patterns between positive and negative items are similar. That is, a high score on a positive item and a low score on a negative item (or vice versa) will evoke similar activity patterns. Along with these results we also found large clusters in the motor cortex. We know these clusters are evoked because participants answered with four digits from each hand on an eight-point scale.

In the second study, we reduced the answer options from eight to two (i.e., applies and does not apply) thereby aiming to reduce motor demands of the task. Contrary to our hypotheses, the searchlight, whole-brain analysis did not return significant results for either the self- or other-imagination-phases. We did however find evidence for coarse- and fine-grained knowledge structures in our ROI analysis within the MPFC with two different ROIs.

During the self-imagination-phase, the Big-5 model was the only significant model within the MPFC ROIs . This model hypothesizes that patterns of activity are similar for items within the same factor and different for items between factors. During the other-imagination-phase, we found evidence for the coarse Big-5 model and additionally for two models that tested finer-grained social knowledge representations. The first similarity model [similarity-1 different] and the explicit similarity model could also explain the patterns of activity. These models hypothesize that patterns of activity are more complex than the Big-5 model and reflect correlations or ratings between individual items.

Overall, across studies we found evidence that both coarser and finer-grained social knowledge structures are represented in cortical patterns of activity during personality learning.

4.4.1 Differences in social aspects of the studies

Interestingly, the results we got for study 1 (personality learning with social interaction) and study 2 (personality learning with no social interaction) were slightly different. Specifically, the searchlight analysis (i.e., the most sensitive analysis) revealed a cluster in the mPFC for the first study during the self-imagination-phase and returned no clusters for the second study. A subsequent more robust ROI analysis for study 2 did reveal complex activity patterns exhibited during personality learning but only during the other-imagination-phase.

We expect that the main factor that drives these differences is the fact that study 1 had social interactions and study 2 did not. The absence of social interactions in the second study might have caused participants to be less engaged in the task, which resulted in less or more noisy cortical activity that could not be detected with the sensitive searchlight analysis. Previous research has found more pronounced activity patterns for social task compared with non-social tasks (Redcay et al., 2010; Rilling et al., 2008).

We surmise that the searchlight only revealed significant activity patterns during the self-imagination-phase for similar reasons. That is, because of the social interactions beforehand participants potentially were more occupied with how they were perceived by the others (Schilbach et al., 2006). This might have caused more distinct patterns of activity during the self-phase compared to the other-phase that were detected with this analysis.

In the second study the searchlight analysis did not reveal any significant clusters. We hypothesize that due to the non-social nature of the task participants were less engaged in the task in general and potentially less occupied with their own personality traits. This is also reflected in

the follow-up ROI analysis. This analysis only revealed coarse granular patterns (i.e., Big-5) during the self-imagination-phase, indicating participants were not using complex fine grained knowledge structures during their self-ratings.

The ROI analysis for the other-imagination-phase did reveal complex knowledge structures (i.e., similarity models). To our knowledge this is the first time such complex structures have been found in cortical activity patterns during personality learning. As hypothesized, we also found evidence for the less complex Big-5 model; this model contains all the coarse information found in the more complex models and has been found before (Thornton & Mitchell, 2017, 2018). Interestingly, we did find evidence for the explicit model. This model hypothesizes that cortical activity patterns reflect participants' idiosyncratic trait word ratings. Recent work already found evidence for idiosyncratic activity to potentially be more prominent than the commonly assumed universal activity patterns (Nakuci et al., 2022).

We did not find evidence for a gradient to exist from ventral to dorsal mPFC as found in other work (Amodio & Frith, 2006; Saxe, 2006; Sul et al., 2015; Wagner et al., 2012). Results were the same for the two ROIs we tested in the second study.

Further technical differences between the two studies and potential limitations in study design for RSA are discussed below.

4.4.2 Potential technical limitations for RSA

In this study, we first reanalyzed data from a previous study. The data from this study was not optimized for RSA but nonetheless gave us promising results with a searchlight analysis. We next collected new data in a similar paradigm but with some changes, aimed at improving the data to be more suited for RSA. However, these changes seemed to have had the contrary effect. In the above section we already discussed the potential impact the (absence) of social interactions could have had on our results. Below, we list in our own judgement of importance, the technical changes we made from one study to the next and their possible (adverse) effects on the results.

4.4.2.1 Timing

In the first study the time distance between trials was 1, 3, or 5 seconds (jittered), whereas the distance between trials for the second study was only 500ms. This small distance between trials could have had a negative effect on the statistical power.

Within the trials there also was a large different in timing. In the first study participants had 4 seconds to imagine/ think about the trait word and subsequently 6 seconds to rate. Compared to the second study where participants had 2.75 seconds for imagination and 1.5 seconds to rate the trait word shows a large difference in timing between the two studies. In general, the BOLD signal had more time to respond fully and return to baseline in the first study. Furthermore, the jittered fixation cross at the end of each trial in the first study would have had a positive effect on the statistical power.

4.4.2.2 Answer Options (Binary vs Scale)

In the original study participants were asked to rate on a scale of 1-8. This allowed for more fine-grained ratings but also evoked a large response in the motor cortex. To avoid this large motor response we opted for a binary response (yes & no, counterbalanced) in the current study. This was effective at reducing the functional motor response. However, we also expect this to have had adverse effects on the functional results. For example, participant engagement could have been decreased by these limited answer options. That is, because they only had to think about whether a trait applied or not they didn't have to engage as much as when thinking exactly how much (or not) this would apply.

4.4.2.3 Number of Stimuli

In study 1 we asked participants to rate both the other and the self on 80 traits total (40 positive and 40 negative). In the current study this was reduced to 40 total traits (20 positive and 20 negative) for both self- and other-ratings. Assuming more stimuli increase power during analysis, there is a trade-off between time in the scanner and amounts of necessary power. By decreasing the number of stimuli by 50% we might have lost too much power to detect effects with less robust analyses e.g., the searchlight, that were effective in the first study. This seems reinforced by the fact that the analysis on the ROI revealed by the searchlight of the original study showed significant model effects in the current study but an independent searchlight did not find significant results in this study.

4.4.2.4 Session Repetition

In the original study there was only one scanning session in which participants rated themselves and others. In the current study we increased the number of sessions to 5 i.e., there was one session,

which had the same stimuli but ordered randomly, that was repeated 5 times. In the analysis we averaged over these five sessions to reduce noise. Repeating sessions to decrease noise still seems like common practice in fMRI analysis, therefore we have put this lowest on our list of potential limitations

4.4.2.5 No Feedback

In the first study participants received feedback after every rating, this feedback phase was absent in the second study. Similar as with the answer options this might have influenced participant engagement.

5. Study 3 – Grid-like encoding in human prefrontal cortex during social navigation

5.1 Introduction and Hypotheses

In the last two studies we found both behavioral and neural evidence for humans using generalizations during personality learning. These generalizations, based on research from personality psychology can distill personality in a coarse manner (i.e., 5 factors where items within each factor are more alike than those between factors) or on a finer granularity where the similarities between all items is considered. Future research might find generalizations on the spectrum between these two extremes. However, for this last study we want to focus on a different generalization, namely the map-like generalizations as computed by grid-cells.

It had already been suggested by the 1950's (Tolman, 1948) that animals might use a map-like structure to make generalizations in their surroundings (e.g., finding novel shortcuts in a maze task). Single cell recordings from rats navigating an arena provided first evidence for these maps in the form of place-cells in the hippocampus (O'Keefe & Dostrovsky, 1971) and grid-cells in the entorhinal cortex (EC) (Hafting et al., 2005). Grid-cells in the human EC during spatial navigation were later also found with direct recordings (Jacobs et al., 2013) as well as non-invasive methods (i.e., fMRI) (Doeller et al., 2010) which opened up research on grid-cells to humans.

Interestingly, the regions implicated in this spatial map (i.e., mPFC, EC, and hippocampus) also support abstract reasoning and decision making (Schuck et al., 2016). In recent years evidence has formed for such non-spatial navigation on a conceptual map within these regions. This conceptual map can, in principle, represent anything that can be projected on two dimensions (Behrens et al., 2018). Studies have found evidence of vastly different maps such as conceptual (Constantinescu et al., 2016), visual (Julian et al., 2018; Nau et al., 2018), olfactory (Bao et al., 2019) and semantic (Viganò et al., 2021; Viganò & Piazza, 2020) and in general suggest a domain-general role for grid-like encoding in the (human) cortex (Bellmund et al., 2018). This would include the social domain for which this grid mechanism was also suggested (Schafer & Schiller, 2018).

So far, two studies have provided first evidence for a social map (Liang et al., 2023; Park et al., 2021). Both studies use a previously found two-dimensional map (i.e., the stereotype content model) (Cuddy et al., 2008; Fiske et al., 2007) with the axes competence & warmth and competence & popularity. Similar to the motivations of our previous works we wanted to explore more detailed representations than the ones studied so far. Since our previous work focused on generalizations derived from the Big-5 and we found evidence for these structures in the cortical activity patterns, we wanted to test whether there was grid-like activity for a personality space based on two dimensions from the Big-5.

5.2 Methods

This experiment was preregistered on the Open Science Framework prior to data collection (https://osf.io/pxs7m). The whole experiment and analysis was conducted accordingly, and deviations and exploratory analyses are mentioned explicitly.

5.2.1 Participants

We recruited 40 participants (27 female, mean age = 25.4 years, SD = 5.0) through online advertisements. We aimed for a sample size similar to (Constantinescu et al., 2016) whose paradigm we closely followed. No participant reported a history of significant medical illness or psychiatric dysfunction. Further inclusion criteria were: 1) age between 18 and 40, 2) Native German speakers and 3) normal or corrected to normal vision. We excluded four participants; three were excluded because they could not finish the behavioral part within the accuracy cut-off, and one because of excessive head movement. This left us with a total of 36 participants who were included in the analysis. The study was conducted in accord with the Declaration of Helsinki and approved by the local research ethics committee (Ethik-Kommission der Ärtzekammer Hamburg, Number: PV5746). All participants gave written consent prior to scanning and were compensated on an hourly basis.

5.2.2 Behavioral Tasks

This experiment was conducted over two days (roughly 6 hours total). The majority of these tasks have been adapted from the first study that looked for grid-cells in conceptual space (Constantinescu et al., 2016) and other recent publications also used very similar task structure within the social domains (Liang et al., 2023; Park et al., 2021). The first day consisted of only behavioral training tasks and the second day was both behavioral training and a neuroimaging task performed in the fMRI scanner. There were a total of nine tasks of which 1 & 2 were only performed on day one, 3-8 were repeated on both days and the 9th task was only performed on the second day. Moreover, on day two, task 7 was changed slightly so it could be performed optimally in the scanner.

In general, we wanted participants to learn about the traits “generous” and “diligent” from six individuals they had never met before. The level to which these traits applied to each stranger

was displayed with the height of two bars (one representing generous and the other diligent), the higher the bar the more this trait applied to this person (Figure 5.1A,B). In order to get participants well acquainted with these bars and how to control their heights as well as their meaning representing personality traits we gave participants multiple tasks to get them familiar with every aspect of this task (Figure 5.1C).

We had two main goals we wanted to accomplish with these tasks. First, for participants to learn the task that was intended for the scanner. That is, learn the meaning of the objects (bars) and their positions on the screen and how to manipulate them. To achieve this we introduce each concept systematically throughout these tasks, with the end goal of teaching participants the personality traits of six strangers in this specific paradigm.

Second, because this was intended to be about social learning but the task itself was rather abstract we added several tasks to increase and test participants' engagement. That is, make sure participants engaged in the task in a meaningful manner, making sure they regarded the stimuli as ratings on personality traits and associated people with this.

All behavioral tasks are explained in brief in the following section with details reserved for the supplementary materials.

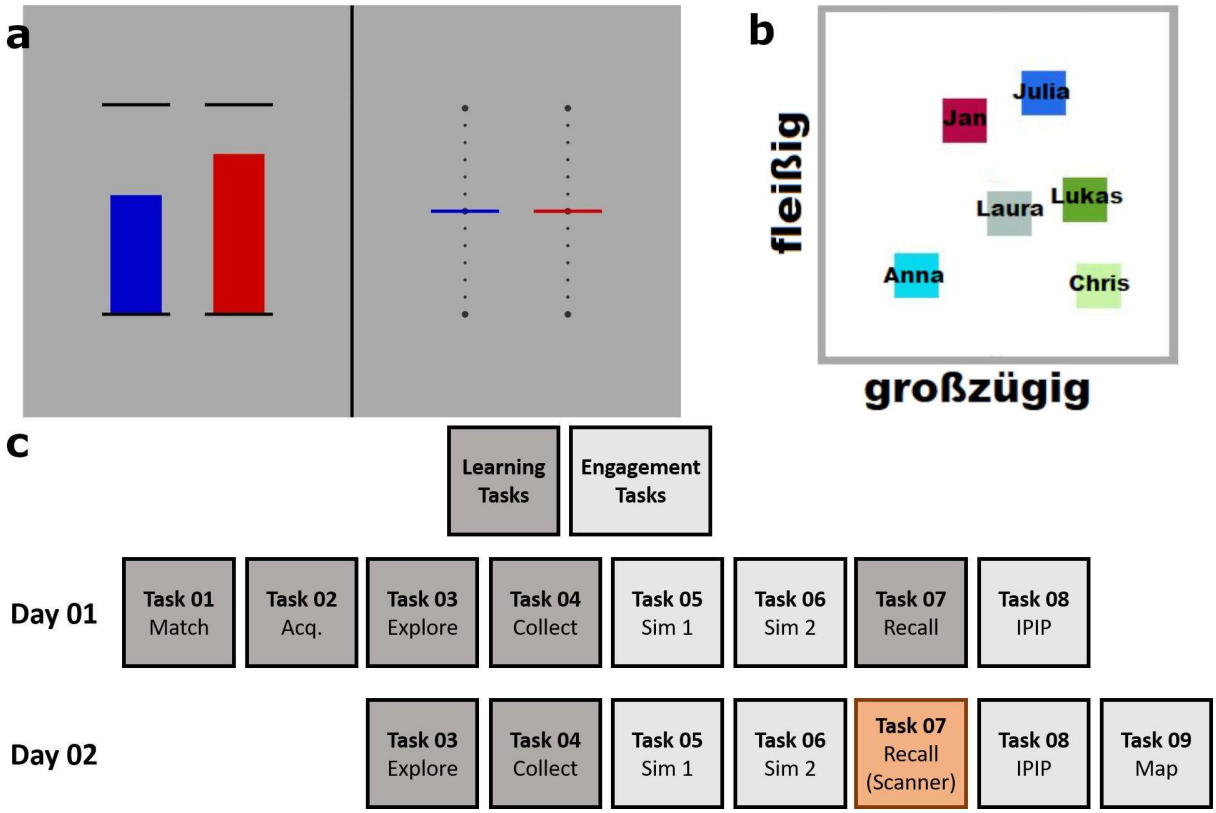


Figure 5.1. Overview of the tasks and general look.

a) The general look for most learning tasks had the sliders on the right and the bars (that represent the personality traits generous and diligent) on the left. Participants could manipulate the bars' heights by changing the slider positions. **b**). The positions of the strangers on the trait map, the axes of the map correspond to the bars' height. **c**) In total there were nine tasks. Tasks 1 and 2 were only performed on day 1 and task 9 only on the second day. Task 7 was performed on the computer in the lab on day 1 and in the scanner on day 2. Tasks can be categorized as learning tasks where participants were either taught how to control the sliders and the bars or the position of the strangers and engagement tasks where we tested if participants engaged with the tasks in the intended manner.

5.2.2.1 Learning tasks

All learning tasks had the same aesthetic throughout the study (Figure 5.1A). These tasks were aimed at getting participants ready for the neuroimaging task. The goal of these tasks was to get participants to learn two personality traits (diligent & generous) of six unknown individuals (hereafter called "strangers"). The personality traits are each represented by a colored bar, of which the height indicates how much of this trait a stranger possesses. There were no time limits on any of the tasks and participants were free to take breaks as they pleased.

In all tasks described below, participants could change the height of the bars by setting the two sliders (each slider corresponded to the same colored bar) above and below the centerline. The settings of the sliders indicate both the movement direction and velocity of the bars. After confirming these slider settings the bars would move for 2 seconds total in the direction and with a velocity based on the sliders' distance from the centerline. Setting the slider to the maximum would move the bars one-third of their total maximum length. The bars could never exceed the maximum or minimum indicated by the horizontal lines.

In the first task [Task 1 - **Match**] participants learned to manipulate the bars (Figure 5.1A, left) using the sliders on the right (Figure 5.1A, right). They were asked to match the height of two transparent bars superimposed on the bars they could move.

After getting participants familiar with manipulating the bars' height we introduce the meaning of the bars i.e., they represent the personality traits diligent and generous. To make it easier for participants to understand this concept we asked them to rate six of their acquaintances on these two traits using the bars in the second task [Task 2 – **Acquaintances**]. That is, they got a prompt for an acquaintance (e.g., a colleague you like) and were then asked to rate how generous and diligent they found this acquaintance. This procedure was repeated for each of six acquaintances (see supplementary table 1 for all prompts). Both tasks 1 and 2 were only performed on day one.

Once participants understood the controls and the personality representations of the bars, they were given the freedom to explore the personality space by changing the heights of both bars at their will. In this personality space six strangers were hidden where the height of the bars corresponded to the coordinates of the personality space (Figure 5.1B). In the third task [Task 3 – **Explore**] participants were asked to find each of these strangers three times (with at least one other stranger in between so they could not cheat and repeat finds).

When participants had found each stranger three times we assumed they were familiar with the strangers' traits. For the fourth task [Task 4 – **Collect**] we asked the participants to change the bar heights from semi-random positions to those of a stranger we indicated. Bar heights were always positioned so participants could reach the stranger within one try (i.e., they were never positioned more than one-third the maximum length away on both axes).

The final learning task on day one was the seventh task [Task 7 – **Recall**]. On the second day, this task was performed in the scanner with inputs from a button box instead of a keyboard. The scanner task is described in detail below and will therefore be omitted here.

5.2.2.2 Engagement tasks

During the engagement tasks we tested whether participants saw the bars as representing personality traits and engaged with them thus. To test whether participants had internalized the strangers' personalities after task 4 we asked them to rate the strangers' similarity (1: 'not similar' – 7: 'very similar') [Task 5 – **Simple Similarities**]. That is, they received each pair of names and were asked to rate how similar the overall personalities of these two strangers were.

In the follow-up task we also asked participants for similarity ratings but now between pairs of strangers [Task 6 – **Complex Similarities**]. That is, participants received two possible pairs of strangers and were asked to decide which of these two pairs were more similar.

To explore whether participants saw the strangers as 'real individuals' (i.e., as multidimensional personalities rather than the 2-dimensional personalities they were presented as) we asked them to rate the strangers on five traits related to the ones they learned about (i.e., traits from the two Big-5 factors agreeableness and conscientiousness) [Task 8 – **IPIP**]. Participants received different traits on day one and two.

The final task, which was only conducted on the second day, revealed the possibility of representing the strangers and acquaintances on a 2-dimensional map. Participants were tasked with dragging both their acquaintances and the strangers to the correct position on the map [Task 9 – **Map**].

5.2.3 Neuroimaging Task [Task 7 - Recall]

After two days of training, where tasks 3 to 6 were repeated before the neuroimaging task, participants performed task 7 in the scanner. Task 7 was the same on both days. The only difference was that on day one it was performed on a regular computer together with the other tasks and subsequently had input from a keyboard and on day two in the scanner with a button box to input answers.

This task was slightly different from the other learning tasks in that participants could not control the movement of the bars anymore, instead the bars moved in pre-defined directions and velocities. On the screen sliders were absent and subsequently the bars were presented in the center of the screen (Figure 5.2).

Trial

At the start of a trial participants were presented with a fixation cross. After the fixation cross the, now familiar, bars (representing diligent and generous) were displayed in the middle of the screen initialized in a random position from the participants' perspective. The bars were motionless for one second. After one second the bars moved in a specific direction and velocity that could be described with a vector on a 2-dimensional plane. After 1.5 seconds of this movement, the bars got covered in a semi-transparent veil which also halted their movement. Participants were instructed to imagine the movement of the bars if they continued with the same direction and velocity for another 1.5 seconds. The passing of these 1.5 seconds was indicated by the border of the veil changing its color from black to white. At this point a trial can develop in two ways (both with equal likelihood 50%, and equal duration of six seconds). It was either an answer trial or an no-answer trial.

During an answer trial, participants were asked to pick one out of three options that were presented on top of the veil simultaneously with the switching of the border color to white. These answer options were always two strangers and an 'empty' option. Participants were tasked with determining at whose personality ratings the bars would have ended at the end of this imagined movement, for this they had 1.5 seconds to choose, after which a fixation cross was presented for .5 second. To make sure answer probabilities stayed equal, correct options were a name option two-thirds of the time and empty one-third of the time.

During a no-answer trial, participants were not presented with any answer options after the 1.5 seconds of imagination had passed but instead got a fixation cross for 2 seconds.

Finally, we added several null-trials to help with statistical robustness. During a null-trial a fixation cross was presented for a jittered duration, which had been calculated previously to ensure an average trial duration of six seconds over all null-trials.

Session

There were a total of 72 actual trials, of which 36 were answer- and 36 were no-answer trials. Furthermore, 14 null-trials were interspersed randomly for a total of 86 trials. With every trial taking on average 6 seconds, this resulted in a total session length of 516 seconds (or approximately 8.5 minutes). Every participant underwent six sessions, which were the same for all participants but presented in pseudorandom order.

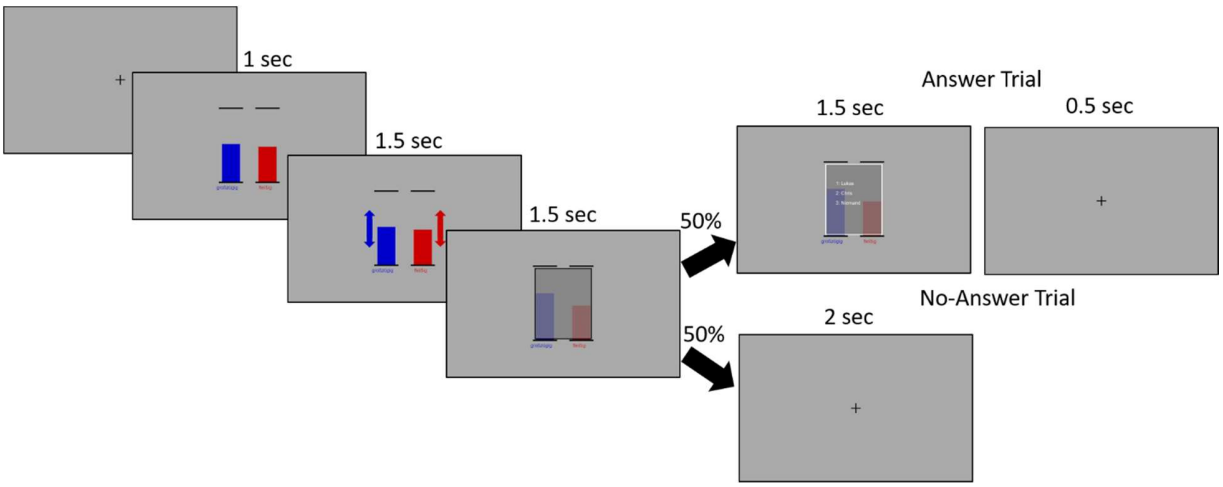


Figure 5.2 Overview of the scanner task

5.2.4 Behavioral data analysis

Behavioral data analysis served two purposes: first to test whether participants learned the tasks and trait space and second to test whether participants engaged with the task in the way we intended (i.e., to perceive the bars as representing personality traits). The analyses are explained in detail in the appendices.

Learning Tasks

We analyzed the learning tasks to make sure participants learned how to manipulate the bars' positions and subsequently learned the positions of the strangers based on their traits. To assess participants' learning (tasks 1, 2, 3, 4, and 7) we mainly compared the performance between day 1 and day 2. Because tasks 1 and 2 were only performed on day one we only plotted these results and visually inspected them for inconsistencies. For tasks 3 and 4 we compared both the number of tries and the total time taken on the first day with those on the second day with two sample t-tests. Furthermore, we visually inspected participants' movement through the trait space to determine if their movement got more localized around the strangers' positions. For task 7 we tested whether participants performed above chance for both days and we compared participants' accuracies between days 1 and 2. We expected an increase in accuracy from day 1 to day 2.

Engagement Tasks

Participants' engagement was tested for tasks 5, 8, and 9. For task 5, where participants rated the strangers' similarities, we performed multidimensional scaling (function 'cmdscale' in Matlab) to project the similarity ratings onto a Euclidean plane. This Euclidean plane was rotated to the best fit with the actual positions of the strangers and visually inspected to determine its fit to the original positions. For task 8 we performed correlation analyses between the ratings on the IPIP items from the same factor as the two trait axes (i.e., the factor agreeableness for generous and conscientiousness for diligent). These correlations were performed for both the strangers and the participants' acquaintances. In a similar vein we performed a correlation analysis for task 9. This correlation was calculated between the actual positions of the strangers and acquaintances (taken from task 2) and where the participants placed them on the map.

5.2.5 MRI Data Acquisition and Pre-processing

We acquired T2*-weighted functional images on a 3 Tesla Siemens Magnetom Prisma with the following parameters: repetition time (TR): 1212ms, echo time (TE): 29ms, FA: 60°, FoV: 224 mm, number of volumes: 432, number of slices: 40, slice thickness: 2 mm with 25% (0.5 mm) gap for a total acquisition time of: 523.2 seconds.

To correct for distortions we acquired a field map with the following parameters: TR: 440ms, TE1: 5.51ms, TE2: 7.97ms, FA: 30°, FoV: 222 mm, slice thickness: 2 mm. We acquire T1-weighted anatomical images using magnetization-prepared rapid gradient echo sequence (MPRAGE) with the following parameters: TE: 2.98ms, TR: 2300ms, FA: 9°, FoV: 256 mm, slice thickness: 1 mm.

Pre-processing

fMRI data were pre-processed using SPM12 software (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) in MATLAB (r2020b). We performed slice time correction, corrected for signal distortions using the field map, and realigned functional scans to the first image using a six-parameter rigid body transformation to correct for motion. The structural image was co-registered to the functional image, and was segmented and normalized to MNI space. After this the realigned functional images were normalized using the transformation parameters from the structural image normalization, and smoothed using an 6 mm full-width at half maximum Gaussian kernel.

5.2.6 Grid-Cell Analysis

Pre-registered analysis

The preregistered analysis largely followed the steps established in Doeller et al., (2010) and Constantinescu et al., (2016). The region of interest (ROI) was defined based on a previous study by Korn et al., (2012) and was modeled as a 10 voxel diameter sphere centered at the coordinates [0, 56, 28]. Data was split in half based on odd / even sessions where both GLM's were applied to one half. All the fMRI analyses presented in this thesis are performed with the Grid Code Analysis Toolbox (GridCAT) (Stangl et al., 2017) with MATLAB r2020b.

GLM one

The first GLM was used to find participants' grid angle in the pre-specified ROI. This was achieved by creating two parametric modulated regressors and six nuisance regressors (to account for motion-related artifacts) that model the movement and imagination phase of the trial (3 seconds). The parametric modulators are Sine and Cosine for the trajectory angle θ_t with a periodicity of 60°, which results in $sin(6 \cdot \theta_t)$ and $cos(6 \cdot \theta_t)$. Consequently, the betas from these regressors should reflect if a region is modulated by hexagonal symmetry. To calculate the personal grid angle φ the betas from sin and cosine are averaged across the voxels in the ROI and these average beta values are put into the following formula $arctan(\beta_{sin} \div \beta_{cos}) \div 6$, which will output the grid angle φ [-30°, 30°].

GLM two

The second GLM, on the second half of the data, functions as an independent consistency check for the personal grid angle calculated in the first GLM. We performed two separate but related analyses to test for the grid-angle consistency. In the first GLM we create one regressor that uses the grid angle calculated in the first half together with 6 nuisance regressors to model motion-related artifacts. The regressor is modeled in the following way $cos(6 \cdot [\theta_t - \varphi])$, where θ_t is the trajectory angle for that specific trial and φ the grid angle calculated in the first GLM. This function assesses whether the grid angle and the trial specific angle align, and if so, assigns high values to the regressor. In the second GLM we create two regressors, one for the aligned trials and the other for the misaligned trials. We determined the trial alignment based on each participants' personal

grid angle that was calculated on the first half of the data. Each trial ±15° the personal grid angle and its multiples of 6 (i.e., 0°, 60°, 120° etc.) was categorized as aligned and the remaining trials as misaligned. For example, if a participant's personal grid angle was 25°, the trials between 10° and 40° (as well as those between 70° and 100° etc.) would be categorized as aligned and trials between 40° and 70° (and its multiples of 6) as misaligned.

For all these analyses we expected the sixfold multiples to be significant because the grid cells pattern the area in a hexagonal pattern. Nevertheless, we also tested multiples of 4, 5, 7 and 8 to make sure our analyses only revealed sixfold patterns.

Regions of Interest

In total we test three ROIs, the first ROI is the one we specified in the preregistration. This is a 10 voxel diameter sphere that is centered on the coordinates [0, 56, 28], which equates to a part of the dorsomedial prefrontal cortex. The second ROI focuses on the ventral part of the mPFC and was taken from other researchers who based this ROI on AAL regions (Liu et al., 2016). The third and final ROI focused on the bilateral entorhinal cortex and was taken from a ROI that was constructed by using ultra high-resolution fMRI (Maass et al., 2015).

5.3 Results

5.3.1 Behavioral

The majority of behavioral results are provided in the appendices. In general participants learned all the tasks well. We specifically saw a significant performance increase for tasks 3 and 4 from day 1 to day 2 (Figure 5.3A,B). Furthermore, visual inspection of participants' movement through the trait space revealed their movement patterns to localize more around the strangers' positions from day 1 to 2 for both task 3 and 4 (Figure 5.3C,D). Participant engagement with the bars as representing the strangers' personality traits was positive. That is, for both tasks 8 and 9 we found positive correlations between the strangers' trait ratings and participants' ratings on related trait words (Figure 5.3E,F, task 8) and their placement of the strangers on a trait map (Figure 5.3G,H, task 9).

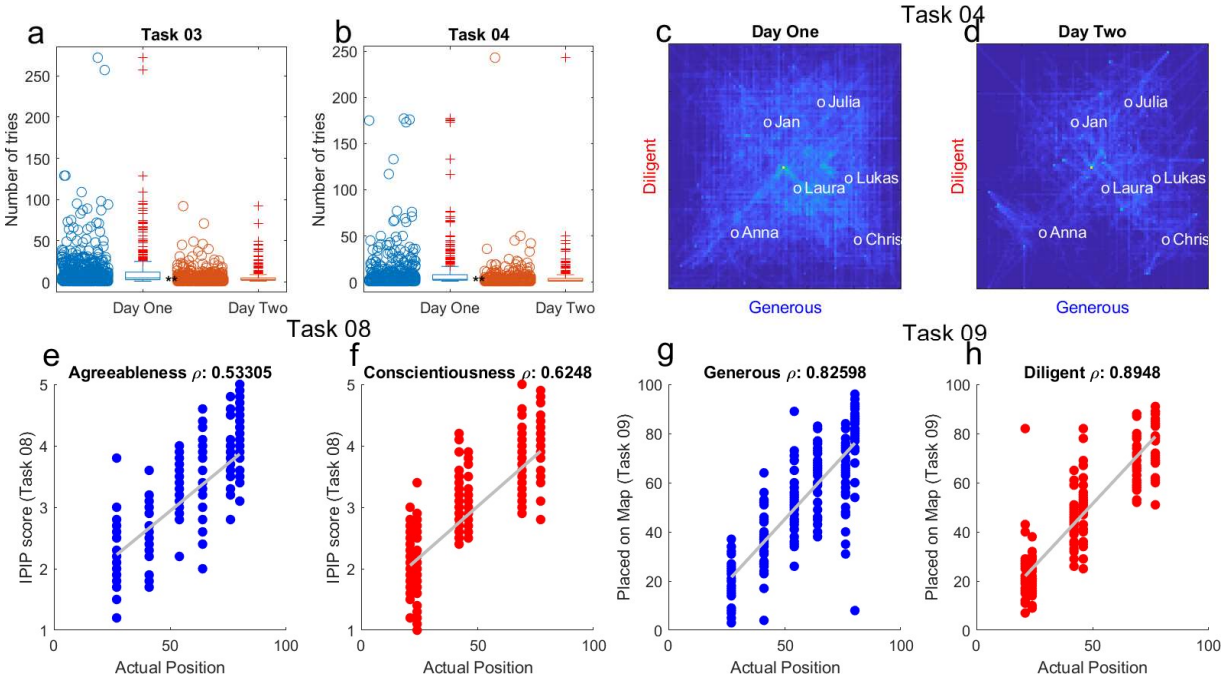


Figure 5.3. Overview of the main behavioral results from the learning and engagement tasks

The top row shows the main results from the learning tasks. **a**) In task 03 [explore] participants got significantly better from day 1 to day 2. **b**) Participants also significantly improved from day 1 to day 2 for task 04 [collect]. To explore participants' movement through the trait space we projected their averaged movements on the map for day 1 (**c**) and day 2 (**d**). Visual inspection shows that participants' movement got more localized around the strangers' positions on day 2. In general we take these results to show that participants learned from day 1 to day 2. The bottom row shows

the main results from the engagement task. **e**) and **f**) show the correlations between participants' IPIP ratings of items from the same factor as generous (agreeableness) and diligent (conscientiousness) with the 6 actual ratings from the strangers on these two traits. The correlations between the ratings for both **e**) and **f**) are satisfactorily, indicating participant engagement. **g**) and **f**) show the correlations between the strangers' actual ratings on generous and diligent and the positions where participants placed them on the map in the final task (09). The correlations for both traits are high, indicating participant engagement.

For task 7 we performed a left-tailed paired samples t-test on participants' accuracies for day 1 (M: .587, SD: .142) and day 2 (M: .698, SD: .134). This revealed significant learning from day 1 to day 2 ($t(35) = -6.7164$, $p<0.001$) (Figure 5.4). Furthermore, two right-tailed paired samples t-test were calculated to test whether participants performed above chance (33%). For both days participants performed above chance, day 1 (M: .587, SD: .142) $t(35) = 10.68$, $p< 0.001$ & day 2 (M: .698, SD: .134) $t(35) = 16.39$, $p< 0.001$, indicating the participants were engaged with the task on both days.

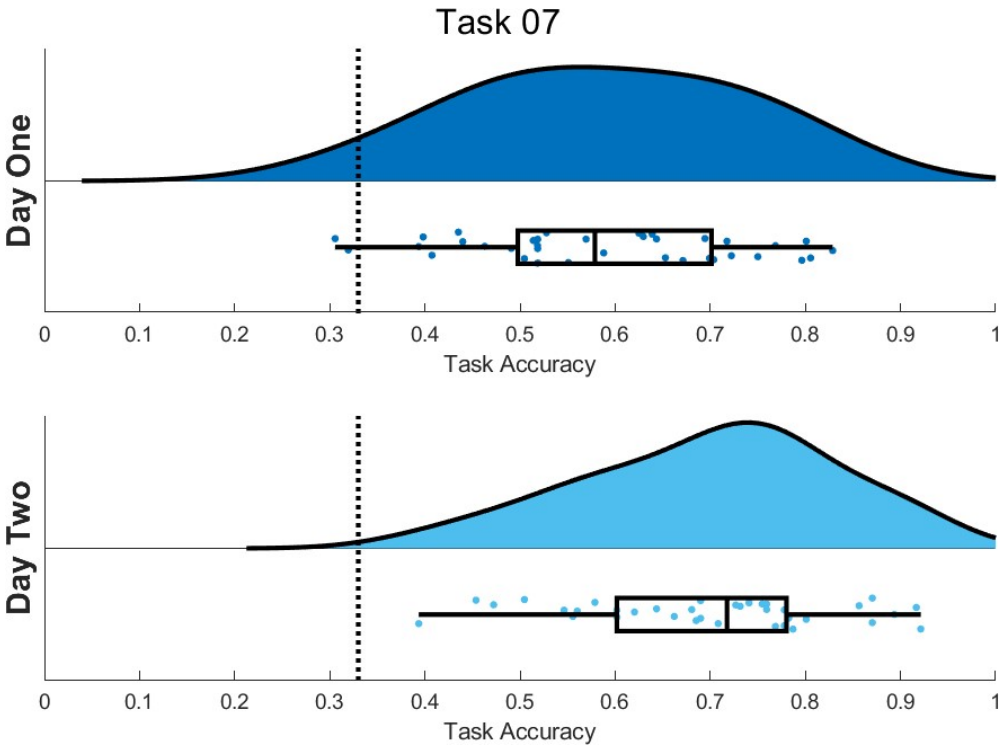


Figure 5.4. Accuracy on task 7 for both days 1 and 2.

During day 1 participants did task 7 on a computer in the lab and on the second day they did this same task in the scanner. Participants' accuracies were significantly above chance (vertical dotted line) for both days and participants' accuracies significantly improved from day 1 to day 2 as tested with a paired samples t-test.

5.3.2 Neural

Having found evidence that participants learned the tasks well and engaged with them in a manner that suggested they learned about the personalities of the strangers we next focussed our attention on the neural analyses. In our main analysis we performed two similar analyses on three distinct ROIs (Figure 5.5A). In them we looked for a hexadirectional modulation of the signal when participants moved through a personality trait space where six strangers were placed based on their levels of generosity and diligence.

In none of the ROIs did we find evidence for a hexadirectional modulation of the signal (Figure 5.5B,C). This indicated that there was no grid-like encoding of the personality trait space for our participants. We next performed a more in depth analysis where we also looked for other modulations of the signal (i.e., other than 6 folds) as a control analysis (Figure 5.6). Similarly to the previous analyses these also did not return any significant results, except for one 5-folded result in the vmPFC.

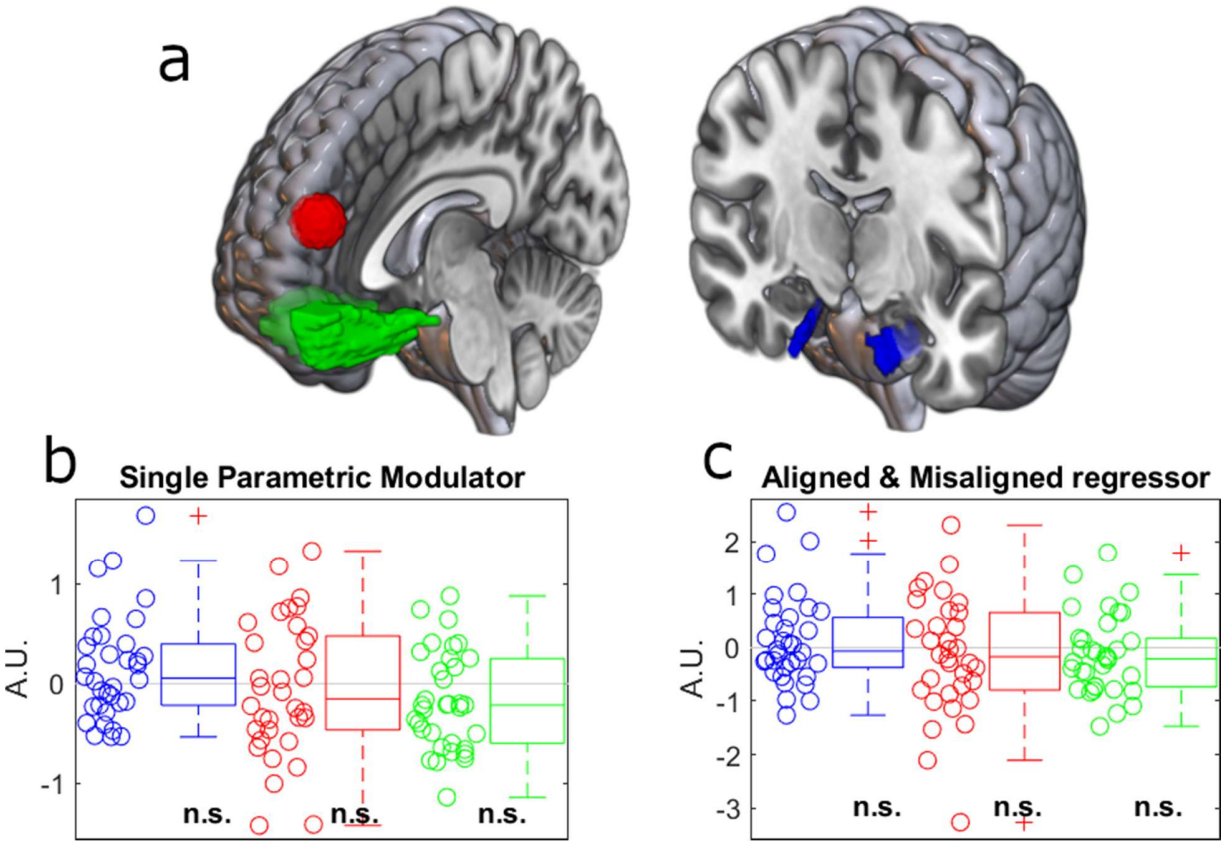


Figure 5.5. Overview of the regions of interest and the main results for each of these regions

a) We tested for grid-like encoding in three different regions of interest. The first region (red) was a sphere in the vmPFC, the second (green) lay in the dmPFC and the third (blue) captured bilateral entorhinal cortex. We tested for

grid consistency with two different analyses both used participants' personal grid angles calculated on the first half of the data. **b**) In the most common analysis we modeled a parametric modulator on the trials based on the personal grid angles. We did not find any signs of grid consistency for any of the ROIs in this analysis. **c**) For the other analysis we sorted the trials based on aligned and misaligned (to the personal grid angle). We also did not find any signs of grid consistency for any of the ROIs for this analysis.

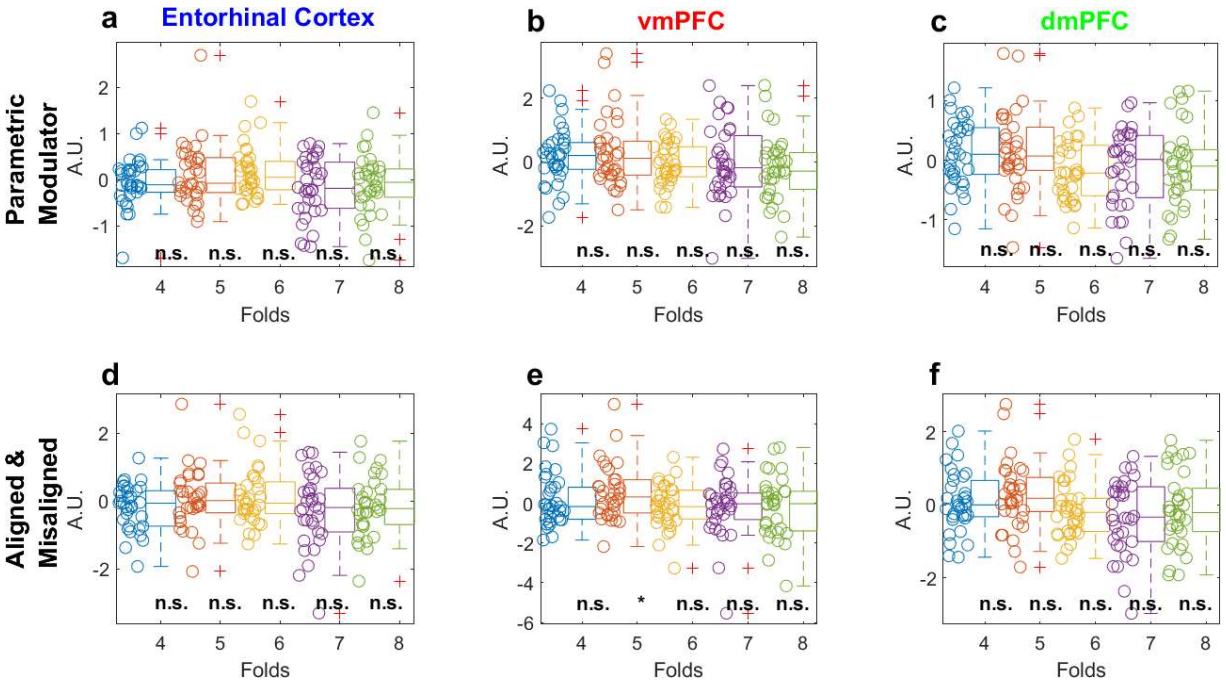


Figure 5.6. Results for all regions of interest on all tested folds

We tested for grid-like encoding on three different ROIs (see figure 5.5a), with two different analyses (first, a parametric modulator based on participants' personal grid angle and second two regressors, one for aligned and the other for misaligned trials). We tested this on five different folds, that is, peaks in the signal. A grid-like signal is only expected for 6 peaks (i.e., hexagonal tiling). We did not find any significant results for any of the analyses but for the 5 folded signal for the aligned and misaligned analysis in the vmPFC (**e**). All alpha values are Bonferroni corrected for the number of tests (i.e., number of folds tested).

5.4 Discussion

In this study we tested whether humans form and traverse a cognitive map of personality traits when learning about strangers' personalities. Over two days we intensively trained participants to get to know our study paradigm (based on (Constantinescu et al., 2016)) and through this paradigm six strangers' generosity and diligence. These two personality traits formed the axes of a two-dimensional conceptual trait space we hypothesized would form in participants' mPFC and entorhinal cortex. We investigated the existence of this trait space by looking for grid-like activity in these regions.

The behavioral analyses revealed participants learned the paradigm and subsequently the strangers' personalities well. Furthermore it showed participants engaged with the task in the way we intended. That is, they learned about the strangers' personality as a whole and not just the positions of the bars representing the two traits (generous and diligent) they were supposed to learn about.

The neural analysis, where we looked for a proxy for grid-like activity in the mPFC and EC, revealed no such coding for any of the ROIs we investigated. We propose our paradigm might be a level of abstraction too deep for the brain to create a cognitive map. When considering other work that investigates grid-like encoding in a social context (Liang et al., 2023; Park et al., 2021) we find that they not only used already established two-factor models but also remained on these factor levels instead of using items within these factors. Our study not only used a five-factor model but traits from within these factors.

5.4.1 Limitations

One of the major limitations in this study could have been the use of a five-factor model. As mentioned above, two other studies within the social domain found grid-like activity for conceptual maps created from two-factor models. Using a five-factor model could complicate results in several ways. The five factors would form a five-dimensional personality plane which one would traverse when learning about others. Such a higher dimensional plane could potentially not be represented by grid cells where evidence so far points to 3-dimensional (Ginosar et al., 2021; R. Hayman et al., 2011; R. M. A. Hayman et al., 2015) as is achievable in spatial navigation. If a five-dimensional plane would be possible to represent with grid cells, interaction effects might skew these planes in

potentially unpredictable ways. A skew of this plane is very likely already happening because these five factors are not exactly orthogonal to each other. This lack of orthogonality means the axes of the maps will not be at 90° which complicates computations. However, we not only used a five-factor model but items from two factors (i.e., generous from agreeableness and diligent from conscientiousness), hypothesizing that a map based on these axes would form during learning. But we cannot be sure that an already existing map (with the formerly mentioned issues) was used instead.

As with any experiment there is a trade-off between external validity and experimental control. In this experiment we leaned towards experimental control and potentially compromised participants' personality learning by being too abstract. Even though we tried to make sure by adding tasks participants did not learn about the strangers in a normal manner.

Learning about others' personality is a continuous process. Using strangers with set personalities gave us a lot of experimental control but one could argue that using participants' acquaintances would offer more external validity. However, the limitations, such as each participant having different maps (because their acquaintances will undoubtedly have different trait ratings) and potentially different scales. In extreme cases these ratings could even change from one day to the next (e.g., when they find out their partner cheated on them).

5.4.2 (Suggested) Future analyses

First, we suggest examining lower level representations that could still show that the brain formed personality representation. For example, investigating whether the Euclidean distance between the strangers' positions on the map and the distance between the activity patterns is similar (using representational similarity analysis (RSA)) would be first evidence of the existence of these personality representations. This analysis could be repeated with participants' placement of the strangers on the map task, potentially revealing participants' erroneous maps of the strangers' positions reflected in their neural activity patterns.

Second, only a subgroup of our participants might have formed a noticeable traits space. We therefore suggest subdividing participants in groups based on their performance in task 7 (e.g., an above and below average performance group) and running the analysis pipeline again for these separate groups hypothesizing that only the above average group will show evidence of a trait space.

Third, because the analysis of grid-like coding from fMRI data is such a novel topic, new and improved analysis pipelines are suggested constantly. One of the most promising analyses uses RSA to explore for grid-like encoding. This analysis foregoes the two GLMs and instead bins the trials (in 6 bins for the main or a different amount of bins for control analyses) and tests whether activity patterns within these bins are similar and between bins dissimilar as a proxy for grid-cells.

5.4.3 Future studies

Future studies should focus on determining whether a personality trait space exists and if so on what level of abstraction. That is, first, find if a personality trait space exists based on the five-factor models. Because grid-cells are based on spatial navigation and such navigation only happens in a maximum of 3 planes a five-dimensional model might not be possible or be distorted. Theoretical work should first prove possible that such computations can be an efficient way of representing higher dimensions. If higher (than 3) dimensional representations are not efficient one might look at maps that get flexibly updated to represent the two or three most important dimensions at that moment in time. Once these factors are sorted one can have a look at the traits within the factors.

Finally, a very interesting study would focus on established representations from acquaintances. These representations will have formed over years and might show strong patterns of activity.

6. Conclusion

In this thesis, I set-out to discover what underlying knowledge structures humans use during learning (with an emphasis on personality learning). I have tried to answer this question on all three levels of Marr's Tri-level hypothesis framework. That is, approach this question from multiple viewpoints to, hopefully, achieve a better understanding of this cognitive process as a whole. The main question, whether humans use knowledge structures, and how exactly they apply them during (social) learning, was the highest level in this framework (i.e., the so-called computational level) and offered the problem that is so elegantly solved in the brain. On the algorithmical level I presented a novel modeling framework that implements knowledge structures in the standard Rescorla-Wagner learning rule. Furthermore, I investigated if and how these knowledge structures are represented in the cortex during learning. I looked at the third level of Marr's framework (implementational) through the lens of grid-cells. The main question here is whether grid-cells also code for a personality trait space like they do for physical space. In brief, this thesis tries to answer a lot of questions about the use of social knowledge structures during social learning in one go. In this final chapter, I will first summarize the results of each chapter. Then I will discuss future directions that could be derived from the current thesis and lastly I will draw the overall conclusions.

6.1 Summary

Social learning and personality learning specifically is of vital importance for healthy social functioning. However, as easy and straightforward this task is for most of us during our everyday existence. It is a vastly complex and ever changing task riddled sometimes with subtleties and other times with generalizations. It is impossible to answer the question of how social learning takes place from only a single vantage point. I therefore try to answer this question from a multitude of perspectives that are based on Marr's framework.

In Chapter 1 ("Introduction") I look at the literature concerning all perspectives I take in this thesis. First, I focus on personality psychology and the generalizations (named knowledge structures throughout the thesis) that have been discovered in this field. These knowledge structures allow humans to distill personality, where the most common of the generalizations (Big-5) captures personality along five dimensions; coarser generalizations on 2 or 3 dimensions also exist, as do

more fine grained ones. I argue that when learning about others' personality humans must make use of all these generalizations depending on the task at hand. After personality psychology I turn my attention towards computational models, specifically, reinforcement learning models and the Rescorla-Wagner learning rule. Over the years these models have proven their worth for explaining details about the brain as well as for studying learning. Together with my co authors, we suggest adding knowledge structures to these models to deepen our understanding of social learning. Next I look at research on the neural processes at play during social learning. With our knowledge from the first study, that humans use different and complex knowledge structures during learning, we look for these structures in cortical activity patterns using multivariate pattern analysis. Research and advantages on multivariate pattern analyses for the work in this thesis is discussed. Finally, I discuss the latest literature on grid-cells, since this is a relatively young field of research I give a brief overview of its largely non-invasive research in human participants.

In Chapter 2 ("Overview of the Methods") my aim was to give a brief introduction to each of the three methods used in this thesis. This was not intended in any way to be an exhaustive tutorial on each of these topics but rather a stepping stone to a basic understanding of each method. My goal was, that after reading about these topics, a motivated person could then go on and fill-in the missing information through their own research. Because all these methods rely quite heavily on code, each subchapter has figures that are made solely using MATLAB code. This code is shared on Github and should help in understanding each of these methods' computational intricacies. Any other teaching materials are shared in their respective sections. The topics are ordered based on the chapters following the current one. First, I look at the Rescorla-Wagner models and our introduction of the knowledge structures. Then I look at how to use representational similarity analysis for behavioral and fMRI data. Finally, I give a brief overview of the analysis used to find a proxy for grid-like activity in fMRI data.

In Chapter 3 ("Study 1 - Finding social knowledge structures using computational models") I discuss the first study of this thesis. For five experiments I discuss the findings from both model-free and those from our modeling procedures. In general I find that participants use different granularities of knowledge structure depending on the task (i.e., available information within the task or the group the other person belongs to). This is the first evidence that humans use complex knowledge structures during social learning. Furthermore, I find that this modeling framework is robust across different sets of stimuli. Showing that this framework has the potential to be used more widely than just for understanding personality learning. All models are tested for robustness

and distinguishability (Wilson & Collins, 2019; L. Zhang et al., 2020). Next, I wondered whether these knowledge structures were also represented in the cortex.

In Chapter 4 ("Study 2 – Patterns of activation during personality representation") I turned towards neural activity patterns. Specifically, those patterns present during social learning in the mPFC. I focussed on the mPFC because along with its activity during social tasks it has also been found to be implicated in accessing conceptual knowledge during decision making, computing decision variables, and prediction errors. That is, all of which was used by participants during the previous study. I report about two experiments which were analyzed for this study. Results are mixed, where on the one hand I find evidence for fine-grained representations in the cortex for the first time but on the other do not find these results consistently. With some evidence for neural activity patterns that match patterns found in behavior, I finally turn towards specific encoding of knowledge structures in the form of grid-cells.

In Chapter 5 ("Study 3 – Grid-like encoding in human prefrontal cortex during social navigation") I look at the coarsest knowledge structure in this thesis (i.e., two dimensions in a trait space). Regions of interest are the mPFC and entorhinal cortex. I report some but not conclusive evidence for grid-like encoding.

6.2 Looking Forward

As with any thesis, I suppose, I have only looked at a small area of research in mine. The studies in this thesis can therefore be improved and expanded in numerous ways to get closer to answering how the brain solves the problem of personality trait learning with respect to the knowledge structures used during this learning process. Below I will illustrate some future directions for each study.

In the first study, I lay-out a novel modeling framework based on the Rescorla-Wagner learning rule. Out of the three studies presented in this thesis, I personally see the most future promises for this study. In reality, I have only presented a very small part of the framework, which already turned into a rather sizable study. Therefore there are some additions to the current models that had to be reserved for future research. These additions of course will allow different questions to be asked. I present some likely contendenders. One could model forgetting over the duration of the experiment by adding a time decay parameter (Kato & Morita, 2016). This would have the

potential of modeling more accurately what participants actually experience during an experiment. Differences in processing positive and negative feedback have been found in previous work (Korn, Sharot, et al., 2014; Sharot et al., 2011), which could be modeled with separate prediction errors for positive and negative feedback. One could then distinguish if learning is different between them. The use of the stereotypical reference point could be elaborated further as research has shown their influence on learning (Hamilton et al., 1990; Jussim et al., 1995; Kang et al., 2021). Exactly when people rely on stereotypical reference points and when they do not might be a fruitful endeavor with positive real life consequences.

From a purely modeling viewpoint the models could be converted to function in a hierarchical Bayesian manner (Ahn et al., 2017) which could increase the power of these models. In general, adopting Bayesian methods together with the Rescorla-Wagner models could provide a best of both worlds scenario (Gershman, 2015). Using a modeling toolbox such as STAN (Carpenter et al., 2017) could provide a convenient way of adopting these methods, moreover, it could benefit the adoption of this framework by other researchers.

From an experimental standpoint, one could increase the external validity by framing the tasks in a more social context such as telling participants they will interact with these people later or even actually interacting with the others such as through a game (Korn et al., 2012).

Use of a psychophysical measure such as eye tracking could prove further evidence for the prediction error. Combining Rescorla-Wagner models together with pupillometry has proven effective in understanding cognitive processes of infants (F. Zhang et al., 2019; F. Zhang & Emberson, 2020). Furthermore, pupil dilation has been successfully linked to rewards processing in noisy tasks (Findling et al., 2019) but also for understanding differences in learning for anxious individuals (Browning et al., 2015; Korn et al., 2017).

A very interesting avenue of research would be to further look into the knowledge structures. Right now the granularity is only used at its two extremes, that is, one coarse and the other fine. Based on previous research (Klein et al., 1992), I expect a gradient to exist between these two and that participants will (unconsciously) use the level of gradient that is appropriate for the task. I hypothesize that a model that uses a free parameter to determine the level of granularity used by the participant to be very beneficial. Furthermore, more research on how these knowledge structures are formed in the first place should still be researched. Research on schemata's (Franklin et al., 2020; Kronenfeld et al., 1978; Mayer & Bower, 1986) might provide a good starting point for understanding the granularities.

In the work presented in this thesis we only look at healthy human beings i.e., those for whom personality learning seems effortless and easy. The true complexity and difficulties attached to personality learning become evident in clinical populations. Important research could therefore be performed in the field of computational psychiatry (Friston et al., 2014; P. R. Montague et al., 2012). As described in a recent publication with my coauthors (Rosenblau et al., 2023), we describe three patient groups who both have marked struggles with personality learning (i.e., autism spectrum disorder ASD, borderline personality disorder BPD, and major depressive disorder MDD) (Decety & Moriguchi, 2007; Gunderson et al., 2018; Millan et al., 2014; Roepke et al., 2013; Yang et al., 2015). We suggest our social-learning framework might be suitable to help therapists to understand their patients but also for patients to understand themselves (Rosenblau et al., 2023).

Finally, even though these models are presented in a personality learning framework they can be applied in many more situations. The strength of these knowledge structures relies on their wide statistical applicability (Roweis & Saul, 2000). That is, they can be applied in many other learning domains, as has already been shown for preference learning (Rosenblau et al., 2021). Any representation that can be specified as connections or relationship structures (Lynn et al., 2020) and their reductions (Hebart et al., 2020) can in principle be applied to our models.

For the second study, which was based on study 1, all new knowledge structures developed in the future for follow-ups of the first study could be tested. That is, the strength of this modeling framework is that all granularities can be tested in the cortex with the analysis pipeline created for study 2. As suggested above, the newly developed intermediate granularity knowledge structures can be directly tested for cortical activity patterns. Furthermore, it would be informative to see the switching between granularities within one experiment. This could potentially also be achieved with a free parameter that models for the current location on the granularity gradient.

In terms of analyses, when calculating the betas in the GLM for each stimulus (i.e., determine the patterns of activity for each stimulus), one could potentially look at using a separate model for each trial (Least Squares Single) instead of estimating all trials within the same model (Least Squares All) to reduce the collinearity (Mumford et al., 2012, 2014). Even though this is usually suggested for interstimulus intervals below 3 seconds, when the ones in the study are around 5 seconds. Similarly, one could change the distance metrics to more reliable metrics. In the current study I used correlation as a distance metric because this was also what was used in the

computational models but different metrics are available and potentially more reliable (Walther et al., 2016).

External validity seems important to get the cortical activity I hypothesized. This is probably coupled to participant engagement. I hypothesize that participants will show more engagement when the task is rooted in actual social activity. An initial experiment that tests whether participant engagement is linked to greater levels of cortical activity should precede any such experiment.

Even though for both studies 1 and 2 I find evidence for more fine grained knowledge structures than have been found so far, I did not find conclusive evidence for a fine-grained knowledge structure in study 3. So far the strongest evidence for a social conceptual space comes from a lower level knowledge structure than the one used in our study. In their study, Park and colleagues (Park et al., 2021) found evidence for grid-like activity for a social hierarchy space based on the two axes competence and popularity. In my study reported here, I have a look at two traits from the Big-5. If the Big-5 is already a higher level knowledge structure than the competence/ popularity one, using its trait words instead of the factors is another level higher. Because these activities are already so miniscule it might not be possible to find activity patterns on these levels of detail yet. That is, methodologies might have to improve first before one can attempt these studies.

An interesting avenue would be to expand to more than two dimensions. Recent research has found three dimensions that fully explain neural representations in humans (Tamir et al., 2016). Testing for grid-like encoding for this third dimension in humans would be very interesting. However, research into this third dimension for spatial navigation found this third dimension to not scale like the other two dimensions for both rats (Grieves et al., 2021; R. Hayman et al., 2011) and bats (Ginosar et al., 2021).

6.3 Final Remarks

Uncovering the statistical structures humans use when learning about others has been actively studied for many years by many people and so have structured representations of neural firing. Understanding human learning takes more than a single thesis but I, nonetheless, have attempted to answer a small part of these questions in this thesis. If the findings in this thesis only contribute a small part to any of these fields I will be satisfied. Most of my hope goes out to the computational framework I developed together with Christoph and Gabriela. This framework together with the neural findings allows for a broader understanding of both healthy and impaired social learning in humans. From a computational psychiatry viewpoint, it offers most direct benefits to pathologies characterized by rigidity such as autism and borderline personality disorder. I also hope that whatever teaching I did together with Christoph, which is largely summarized in chapter 2, will be of benefit to others so they can go on and make their own small contribution to these fields.

References

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*(0), 24. https://doi.org/10.1162/CPSY_a_00002

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277. https://doi.org/10.1038/nrn1884

Ashton, M. C. (2018). *Individual differences and personality* (Third Edition). Academic Press, an imprint of Elsevier Ltd.

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bao, X., Gjorgieva, E., Shanahan, L. K., Howard, J. D., Kahnt, T., & Gottfried, J. A. (2019). Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. *Neuron*, *102*(5), 1066-1075.e5. https://doi.org/10.1016/j.neuron.2019.03.034

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J., Deuker, L., Navarro Schröder, T., & Doeller, C. F. (2016). Grid-cell representations in mental simulation. *eLife*, *5*, e17089. https://doi.org/10.7554/eLife.17089

Bellmund, J., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415), eaat6766. https://doi.org/10.1126/science.aat6766

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, *12*(1), 4081. https://doi.org/10.1038/s41467-021-24368-2

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using Grid Cells for Navigation. *Neuron*, *87*(3), 507–520. https://doi.org/10.1016/j.neuron.2015.07.006

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational Similarity Mapping of Distributional Semantics in Left Inferior Frontal, Middle Temporal, and Motor Cortex. *Cerebral Cortex*, cercor;bhw379v1. https://doi.org/10.1093/cercor/bhw379

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). *Stan*: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, *38*(4), 476–506. https://doi.org/10.1037/h0054116

Collins, A. G. E., & Shenhav, A. (2022). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, *47*(1), 104–118. https://doi.org/10.1038/s41386-021-01126-y

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In *Advances in Experimental Social Psychology* (Vol. 40, pp. 61–149). Elsevier. https://doi.org/10.1016/S0065-2601(07)00002-0

Cushman, F. (2024). Computational Social Psychology. *Annual Review of Psychology*, *75*(1), annurev-psych-021323-040420. https://doi.org/10.1146/annurev-psych-021323-040420

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Daw, N. D., & Tobler, P. N. (2014). Value Learning through Reinforcement. In *Neuroeconomics* (pp. 283–298). Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00015-2

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

Decety, J., & Moriguchi, Y. (2007). The empathic brain and its dysfunction in psychiatric populations: Implications for intervention across different clinical conditions. *BioPsychoSocial Medicine*, *1*(1), 22. https://doi.org/10.1186/1751-0759-1-22

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(11), 1611–1618. https://doi.org/10.1038/nn1575

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634. https://doi.org/10.1093/scan/nsw171

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational Similarity Analyses. In *Handbook of Behavioral Neuroscience* (Vol. 28, pp. 509–525). Elsevier. https://doi.org/10.1016/B978-0-12-812028-6.00027-6

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661. https://doi.org/10.1038/nature08704

Dunne, S., & O'Doherty, J. P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology*, *23*(3), 387–392. https://doi.org/10.1016/j.conb.2013.02.007

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Etzel, J. A., & Braver, T. S. (2013). MVPA Permutation Schemes: Permutation Testing in the Land of Cross-Validation. *2013 International Workshop on Pattern Recognition in Neuroimaging*, 140–143. https://doi.org/10.1109/PRNI.2013.44

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational Substrates of Social Value in Interpersonal Collaboration. *Journal of Neuroscience*, *35*(21), 8170–8180. https://doi.org/10.1523/JNEUROSCI.4775-14.2015

Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. HTTP://sfx-49gbv.hosted.exlibrisgroup.com/sfx_subhh?sid=google&auinit=S&aulast=Farrell&title=

Computational%20modeling%20of%20cognition%20and%20behavior&genre=book&isbn=1108548245&date=2018

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7). https://doi.org/10.1073/pnas.1715227115

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Forstmann, B. U., & Wagenmakers, E.-J. (Eds.). (2015). *An Introduction to Model-Based Cognitive Neuroscience*. Springer New York. https://doi.org/10.1007/978-1-4939-2236-9

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured Event Memory: A neuro-symbolic model of event cognition. *Psychological Review*, *127*(3), 327–361. https://doi.org/10.1037/rev0000177

Freeman, J. B., Stolier, R. M., Brooks, J. A., & Stillerman, B. S. (2018). The neural representational geometry of social perception. *Current Opinion in Psychology*, *24*, 83–91. https://doi.org/10.1016/j.copsyc.2018.10.003

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 671–678. https://doi.org/10.1098/rstb.2006.2003

Frolichs, K. M. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications*, *13*(1), 6205. https://doi.org/10.1038/s41467-022-33418-2

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. J., & Dolan, R. J. (2015). Learning-Induced Plasticity in Medial Prefrontal Cortex Predicts Preference Malleability. *Neuron*, *85*(2), 418–428. https://doi.org/10.1016/j.neuron.2014.12.033

Garvert, M. M., Saanum, T., Schulz, E., Schuck, N. W., & Doeller, C. F. (2023). Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nature Neuroscience*, *26*(4), 615–626. https://doi.org/10.1038/s41593-023-01283-x

Gershman, S. J. (2015). A Unifying Probabilistic View of Associative Learning. *PLOS Computational Biology*, *11*(11), e1004567. https://doi.org/10.1371/journal.pcbi.1004567

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Ginosar, G., Aljadeff, J., Burak, Y., Sompolinsky, H., Las, L., & Ulanovsky, N. (2021). Locally ordered representation of 3D space in the entorhinal cortex. *Nature*, *596*(7872), 404–409. https://doi.org/10.1038/s41586-021-03783-x

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Goldberg, L. R. (1990). An alternative 'description of personality': The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Grieves, R. M., Jedidi-Ayoub, S., Mishchanchuk, K., Liu, A., Renaudineau, S., Duvelle, É., & Jeffery, K. J. (2021). Irregular distribution of grid cell firing fields in rats exploring a 3D volumetric space. *Nature Neuroscience*, *24*(11), 1567–1573. https://doi.org/10.1038/s41593-021-00907-4

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nature Reviews Disease Primers*, *4*(1), 18029. https://doi.org/10.1038/nrdp.2018.29

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. https://doi.org/10.1038/nature03721

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-Based Expectancies: Effects on Information Processing and Social Behavior. *Journal of Social Issues*, *46*(2), 35–60. https://doi.org/10.1111/j.1540-4560.1990.tb01922.x

Hartley, C. A., & Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences*, *5*, 108–115. https://doi.org/10.1016/j.cobeha.2015.09.004

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine All the People: How the Brain Creates and Uses Personality Models to Predict Behavior. *Cerebral Cortex*, *24*(8), 1979–1987. https://doi.org/10.1093/cercor/bht042

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hayman, R. M. A., Casali, G., Wilson, J. J., & Jeffery, K. J. (2015). Grid cells on steeply sloping terrain: Evidence for planar rather than volumetric encoding. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00925

Hayman, R., Verriotis, M. A., Jovalekic, A., Fenton, A. A., & Jeffery, K. J. (2011). Anisotropic encoding of three-dimensional space by place cells and grid cells. *Nature Neuroscience*, *14*(9), 1182–1188. https://doi.org/10.1038/nn.2892

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications*, *7*(1), 12722. https://doi.org/10.1038/ncomms12722

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X.-X., Suthana, N., Sperling, M. R., Sharan, A. D., Fried, I., & Kahana, M. J. (2013). Direct recordings of grid-

like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*(9), 1188–1190. https://doi.org/10.1038/nn.3466

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017a). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017b). Social learning through prediction error in the brain. *Npj Science of Learning*, *2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2

Jolly, E., & Chang, L. J. (2019). The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Topics in Cognitive Science*, *11*(2), 433–454. https://doi.org/10.1111/tops.12404

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., Voss, H. U., Ballon, D. J., & Casey, B. J. (2011). Behavioral and Neural Properties of Social Reinforcement Learning. *Journal of Neuroscience*, *31*(37), 13039–13045. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

Julian, J. B., Keinath, A. T., Frazzetta, G., & Epstein, R. A. (2018). Human entorhinal cortex represents visual space using a boundary-anchored grid. *Nature Neuroscience*, *21*(2), 191–194. https://doi.org/10.1038/s41593-017-0049-1

Jussim, L., Nelson, T. E., Manis, M., & Soffin, S. (1995). Prejudice, stereotypes, and labeling effects: Sources of bias in person perception. *Journal of Personality and Social Psychology*, *68*(2), 228–246. https://doi.org/10.1037/0022-3514.68.2.228

Kahnt, T., & Tobler, P. N. (2016). Dopamine regulates stimulus generalization in the human hippocampus. *eLife*, *5*, e12678. https://doi.org/10.7554/eLife.12678

Kang, P., Burke, C. J., Tobler, P. N., & Hein, G. (2021). Why We Learn Less from Observing Outgroups. *The Journal of Neuroscience*, *41*(1), 144–152. https://doi.org/10.1523/JNEUROSCI.0926-20.2020

Kanske, P. (2018). The social mind: Disentangling affective and cognitive routes to understanding others. *Interdisciplinary Science Reviews*, *43*(2), 115–124. https://doi.org/10.1080/03080188.2018.1453243

Kato, A., & Morita, K. (2016). Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. *PLOS Computational Biology*, *12*(10), e1005145. https://doi.org/10.1371/journal.pcbi.1005145

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology*, *63*(5), 739–753. https://doi.org/10.1037/0022-3514.63.5.739

Klein-Flügge, M. C., Bongioanni, A., & Rushworth, M. F. S. (2022). Medial and orbital frontal cortex in decision-making and flexible behavior. *Neuron*, *110*(17), 2743–2770. https://doi.org/10.1016/j.neuron.2022.05.022

Korn, C. W., Fan, Y., Zhang, K., Wang, C., Han, S., & Heekeren, H. R. (2014). Cultural influences on social feedback processing of character traits. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00192

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592. https://doi.org/10.1017/S0033291713001074

Korn, C. W., Staib, M., Tzovara, A., Castegnetti, G., & Bach, D. R. (2017). A pupil size response model to assess fear learning. *Psychophysiology*, *54*(3), 330–343. https://doi.org/10.1111/psyp.12801

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kronenfeld, D. B., Schank, R. C., & Abelson, R. P. (1978). Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. *Language*, *54*(3), 779. https://doi.org/10.2307/412850

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/j.neuron.2009.07.030

Lau, B., & Glimcher, P. W. (2005). DYNAMIC RESPONSE-BY-RESPONSE MODELS OF MATCHING BEHAVIOR IN RHESUS MONKEYS. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. https://doi.org/10.1901/jeab.2005.110-04

Lee, D., Seo, H., & Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience*, *35*(1), 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512

Lee, K., & Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Liang, Z., Wu, S., Wu, J., Wang, W., Qin, S., & Liu, C. (2023). *Social navigation: Distance and grid-like codes support navigation of abstract social space in human brain* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.05.12.538784

Liu, Z.-X., Grady, C., & Moscovitch, M. (2016). Effects of Prior-Knowledge on Brain Activation and Connectivity During Associative Memory Encoding. *Cerebral Cortex*, bhw047. https://doi.org/10.1093/cercor/bhw047

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011

Lockwood, P. L., & Klein-Flügge, M. C. (2020). Computational modelling of social cognition and behaviour—A reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, nsaa040. https://doi.org/10.1093/scan/nsaa040

Love, B. C. (2015). The Algorithmic Level Is the Bridge Between Computation and Brain. *Topics in Cognitive Science*, *7*(2), 230–242. https://doi.org/10.1111/tops.12131

Lynn, C. W., Kahn, A. E., Nyema, N., & Bassett, D. S. (2020). Abstract representations of events arise from mental errors in learning and memory. *Nature Communications*, *11*(1), 2313. https://doi.org/10.1038/s41467-020-15146-7

Maass, A., Berron, D., Libby, L. A., Ranganath, C., & Düzel, E. (2015). Functional subregions of the human entorhinal cortex. *eLife*, *4*, e06426. https://doi.org/10.7554/eLife.06426

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46. https://doi.org/10.1038/s41467-019-13930-8

Mayer, J. D., & Bower, G. H. (1986). Learning and memory for personality prototypes. *Journal of Personality and Social Psychology*, *51*(3), 473–492. https://doi.org/10.1037/0022-3514.51.3.473

McCrae, R., & Costa Jr., P. T. (2008). The five-factor theory of personality. In *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). The Guildford Press.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., Costa, P. T., de Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., Marušić, I., Bratko, D., Caprara, G. V., Barbaranelli, C., Chae, J.-H., & Piedmont, R. L. (1999). Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology*, *35*(2), 466–477. https://doi.org/10.1037/0012-1649.35.2.466

Meyer, M. L., Davachi, L., Ochsner, K. N., & Lieberman, M. D. (2019). Evidence That Default Network Connectivity During Rest Consolidates Social Information. *Cerebral Cortex*, *29*(5), 1910–1920. https://doi.org/10.1093/cercor/bhy071

Meyer, M. L., & Lieberman, M. D. (2018). Why People Are Always Thinking about Themselves: Medial Prefrontal Cortex Activity during Rest Primes Self-referential Processing. *Journal of Cognitive Neuroscience*, *30*(5), 714–721. https://doi.org/10.1162/jocn_a_01232

Millan, M. J., Fone, K., Steckler, T., & Horan, W. P. (2014). Negative symptoms of schizophrenia: Clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. *European Neuropsychopharmacology*, *24*(5), 645–692. https://doi.org/10.1016/j.euroneuro.2014.03.008

Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1123–1136. https://doi.org/10.1098/rstb.2002.1099

Mitchell, J. P. (2006). Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research*, *1079*(1), 66–75. https://doi.org/10.1016/j.brainres.2005.12.113

Mitchell, J. P. (2008). Contributions of Functional Neuroimaging to the Study of Social Cognition. *Current Directions in Psychological Science*, *17*(2), 142–146. https://doi.org/10.1111/j.1467-8721.2008.00564.x

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The Link between Social Cognition and Self-referential Thought in the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315. https://doi.org/10.1162/0898929055002418

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-Specific Effects of Social Cognition on the Neural Correlates of Subsequent Memory. *The Journal of Neuroscience*, *24*(21), 4912–4917. https://doi.org/10.1523/JNEUROSCI.0481-04.2004

Montague, P., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLOS Biology*, *18*(12), e3001028. https://doi.org/10.1371/journal.pbio.3001028

Nakuci, J., Yeon, J., Xue, K., Kim, J.-H., Kim, S.-P., & Rahnev, D. (2022). *Quantifying the contribution of subject and group factors in brain activation* [Preprint]. Neuroscience. https://doi.org/10.1101/2022.08.01.502338

Nau, M., Navarro Schröder, T., Bellmund, J. L. S., & Doeller, C. F. (2018). Hexadirectional coding of visual space in human entorhinal cortex. *Nature Neuroscience*, *21*(2), 188–190. https://doi.org/10.1038/s41593-017-0050-8

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73. https://doi.org/10.1016/j.cobeha.2016.04.005

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175. https://doi.org/10.1016/0006-8993(71)90358-1

O'Keefe, J., & Nadel, L. (1978). The Hippocampus as a Cognitive Map. *Oxford Univ. Press*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Park, S. A., Miller, D. S., & Boorman, E. D. (2021). Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, *24*(9), 1292–1301. https://doi.org/10.1038/s41593-021-00916-3

Park, S. A., Sestito, M., Boorman, E. D., & Dreher, J.-C. (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications*, *10*(1), 5287. https://doi.org/10.1038/s41467-019-12937-5

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*(5), 0072. https://doi.org/10.1038/s41562-017-0072

Parkinson, C., Liu, S., & Wheatley, T. (2014). A Common Cortical Metric for Spatial, Temporal, and Social Distance. *The Journal of Neuroscience*, *34*(5), 1979–1987. https://doi.org/10.1523/JNEUROSCI.2159-13.2014

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of Neurosciences*, *17*(3). https://doi.org/10.5214/ans.0972-7531.1017309

Peabody, D., & De Raad, B. (2002). The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology*, *83*(4), 983–997. https://doi.org/10.1037/0022-3514.83.4.983

Peer, M., Hayman, M., Tamir, B., & Arzy, S. (2021). Brain Coding of Social Network Structure. *The Journal of Neuroscience*, *41*(22), 4897–4909. https://doi.org/10.1523/JNEUROSCI.2641-20.2021

Popal, H., Wang, Y., & Olson, I. R. (2019). A Guide to Representational Similarity Analysis for Social Neuroscience. *Social Cognitive and Affective Neuroscience*, *14*(11), 1243–1253. https://doi.org/10.1093/scan/nsz099

Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D. E., & Saxe, R. (2010). Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. *NeuroImage*, *50*(4), 1639–1647. https://doi.org/10.1016/j.neuroimage.2010.01.052

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton-Century-Crofts.

Riberto, M., Paz, R., Pobric, G., & Talmi, D. (2022). The Neural Representations of Emotional Experiences Are More Similar Than Those of Neutral Experiences. *The Journal of Neuroscience*, *42*(13), 2772–2785. https://doi.org/10.1523/JNEUROSCI.1490-21.2022

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *NeuroImage*, *41*(4), 1447–1461. https://doi.org/10.1016/j.neuroimage.2008.03.044

Roepke, S., Vater, A., Preißler, S., Heekeren, H. R., & Dziobek, I. (2013). Social cognition in borderline personality disorder. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00195

Rosenblau, G., Frolichs, K., & Korn, C. W. (2023). A neuro-computational social learning framework to facilitate transdiagnostic classification and treatment across psychiatric disorders. *Neuroscience & Biobehavioral Reviews*, *149*, 105181. https://doi.org/10.1016/j.neubiorev.2023.105181

Rosenblau, G., Korn, C. W., Dutton, A., Lee, D., & Pelphrey, K. A. (2021). Neurocognitive Mechanisms of Social Inferences in Typical and Autistic Adolescents. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(8), 782–791. https://doi.org/10.1016/j.bpsc.2020.07.002

Rosenblau, G., Korn, C. W., & Pelphrey, K. A. (2018). A Computational Account of Optimizing Social Predictions Reveals That Adolescents Are Conservative Learners in Social Contexts. *The Journal of Neuroscience*, *38*(4), 974–988. https://doi.org/10.1523/JNEUROSCI.1044-17.2017

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Saucier, G. (2009). Recurrent Personality Dimensions in Inclusive Lexical Studies: Indications for a Big Six Structure. *Journal of Personality*, *77*(5), 1577–1614. https://doi.org/10.1111/j.1467-6494.2009.00593.x

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235–239. https://doi.org/10.1016/j.conb.2006.03.001

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding Other Minds: Linking Developmental Psychology and Functional Neuroimaging. *Annual Review of Psychology*, *55*(1), 87–124. https://doi.org/10.1146/annurev.psych.55.090902.142044

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind. *Psychological Science*, *17*(8), 692–699. https://doi.org/10.1111/j.1467-9280.2006.01768.x

Schafer, M., & Schiller, D. (2018). Navigating Social Space. *Neuron*, *100*(2), 476–489. https://doi.org/10.1016/j.neuron.2018.10.006

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730. https://doi.org/10.1016/j.neuropsychologia.2005.07.017

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P.,

Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. https://doi.org/10.1038/nn.2949

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Skerry, A. E., & Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, *25*(15), 1945–1954. https://doi.org/10.1016/j.cub.2015.06.009

Soto, C. J., Kronauer, A., & Liang, J. K. (2015). Five-Factor Model of Personality. In S. K. Whitbourne (Ed.), *The Encyclopedia of Adulthood and Aging* (1st ed., pp. 1–5). Wiley. https://doi.org/10.1002/9781118521373.wbeaa014

Stangl, M., Shine, J., & Wolbers, T. (2017). The GridCAT: A Toolbox for Automated Analysis of Human Grid Cell Codes in fMRI. *Frontiers in Neuroinformatics*, *11*, 47. https://doi.org/10.3389/fninf.2017.00047

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour*, *4*(4), 361–371. https://doi.org/10.1038/s41562-019-0800-6

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching Behavior and the Representation of Value in the Parietal Cortex. *Science*, *304*(5678), 1782–1787. https://doi.org/10.1126/science.1094765

Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences*, *112*(25), 7851–7856. https://doi.org/10.1073/pnas.1423895112

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, *22*(3), 201–212. https://doi.org/10.1016/j.tics.2017.12.005

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194–199. https://doi.org/10.1073/pnas.1511905112

Theves, S., Neville, D. A., Fernández, G., & Doeller, C. F. (2021). Learning and Representation of Hierarchical Concepts in Hippocampus and Prefrontal Cortex. *The Journal of Neuroscience*, *41*(36), 7675–7686. https://doi.org/10.1523/JNEUROSCI.0657-21.2021

Thornton, M. A., & Mitchell, J. P. (2017). Consistent Neural Activity Patterns Represent Personally Familiar People. *Journal of Cognitive Neuroscience*, *29*(9), 1583–1594. https://doi.org/10.1162/jocn_a_01151

Thornton, M. A., & Mitchell, J. P. (2018). Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing. *Cerebral Cortex*, *28*(10), 3505–3520. https://doi.org/10.1093/cercor/bhx216

Thornton, M. A., Rmus, M., Vyas, A. D., & Tamir, D. I. (2023). Transition dynamics shape mental state concepts. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001405

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. *Nature Communications*, *10*(1), 2117. https://doi.org/10.1038/s41467-019-10083-6

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019). The brain represents people as the mental states they habitually experience. *Nature Communications*, *10*(1), 2291. https://doi.org/10.1038/s41467-019-10309-7

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tupes, E. C., & Christal, R. E. (1992). Recurrent Personality Factors Based on Trait Ratings. *Journal of Personality*, *60*(2), 225–251. https://doi.org/10.1111/j.1467-6494.1992.tb00973.x

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Viganò, S., & Piazza, M. (2020). Distance and Direction Codes Underlie Navigation of a Novel Semantic Space in the Human Brain. *The Journal of Neuroscience*, *40*(13), 2727–2736. https://doi.org/10.1523/JNEUROSCI.1849-19.2020

Viganò, S., Rubino, V., Soccio, A. D., Buiatti, M., & Piazza, M. (2021). Grid-like and distance codes for representing word meaning in the human brain. *NeuroImage*, *232*, 117876. https://doi.org/10.1016/j.neuroimage.2021.117876

Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *WIREs Cognitive Science*, *3*(4), 451–470. https://doi.org/10.1002/wcs.1183

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Weaverdyck, M. E., Thornton, M. A., & Tamir, D. I. (2021). The representational structure of mental states generalizes across target people and stimulus modalities. *NeuroImage*, *238*, 118258. https://doi.org/10.1016/j.neuroimage.2021.118258

Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In *Handbook of personality psychology* (pp. 737–765). Academic Press.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2020). Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Computational Biology*, *16*(9), e1008149. https://doi.org/10.1371/journal.pcbi.1008149

Yang, D. Y.-J., Rosenblau, G., Keifer, C., & Pelphrey, K. A. (2015). An integrative neural model of social perception, action observation, and theory of mind. *Neuroscience & Biobehavioral Reviews*, *51*, 263–275. https://doi.org/10.1016/j.neubiorev.2015.01.020

Zaki, J., Kallman, S., Wimmer, G. E., Ochsner, K., & Shohamy, D. (2016). Social Cognition as Reinforcement Learning: Feedback Modulates Emotion Inference. *Journal of Cognitive Neuroscience*, *28*(9), 1270–1282. https://doi.org/10.1162/jocn_a_00978

Zhang, F., & Emberson, L. L. (2020). Using pupillometry to investigate predictive processes in infancy. *Infancy*, *25*(6), 758–780. https://doi.org/10.1111/infa.12358

Zhang, F., Jaffe-Dax, S., Wilson, R. C., & Emberson, L. L. (2019). Prediction in infants and adults: A pupillometry study. *Developmental Science*, *22*(4), e12780. https://doi.org/10.1111/desc.12780

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

Appendices

Study 1

Item number	German trait word used	Approximate English translation
Factor 1: neuroticism – positive traits		
1	gelassen	composed
2	locker	easy-going
3	selbstständig	self-reliant
4	souverän	confident
Factor 1: neuroticism – negative traits		
5	ängstlich	anxious
6	feige	cowardly
7	launisch	moody
8	unentschlossen	indecisive
9	unsicher	insecure
10	wehleidig	whiny
Factor 2: extraversion – positive traits		
11	enthusiastisch	enthusiastic
12	lebenslustig	fun-loving
13	schlagfertig	articulate
14	tatkräftig	dynamic
Factor 2: extraversion – negative traits		
15	humorlos	humorless
16	kalt	cold-hearted
17	scheu	unassertive
18	unnahbar	inapproachable
Factor 3: openness to experience – positive traits		
19	kreativ	creative

20	offen	open-minded
21	spontan	spontaneous
22	tolerant	tolerant
23	vielseitig	versatile
24	wissbegierig	inquisitive
Factor 3: openness to experience – negative traits		
25	bieder	overly conservative
26	engstirnig	narrow-minded
27	träge	lazy
28	voreingenommen	biased
Factor 4: agreeableness – positive traits		
29	einfühlsam	empathetic
30	freundlich	friendly
31	gesellig	sociable
32	großzügig	generous
33	hilfsbereit	helpful
34	höflich	polite
35	respektvoll	Respectful
36	vertrauenswürdig	trustworthy
37	zuverlässig	reliable
Factor 4: agreeableness – negative traits		
38	aggressiv	aggressive
39	arrogant	arrogant
40	egoistisch	selfish
41	eitel	conceited
42	gehässig	spiteful
43	großmäulig	loud-mouthed
44	hinterhältig	conniving
45	rücksichtslos	inconsiderate

46	stur	stubborn
47	unsympathisch	unpleasant
Factor 5: conscientiousness – positive traits		
48	aufrichtig	honest
49	bescheiden	modest
50	diszipliniert	organized
51	effizient	efficient
52	fleißig	hard-working
53	kompetent	competent
54	ordentlich	tidy
Factor 5: conscientiousness – negative traits		
55	chaotisch	chaotic
56	inkonsequent	inconsistent
57	leichtsinnig	foolhardy
58	pedantisch	pedantic
59	unpünktlich	tardy
60	zwanghaft	obsessive

Table 9.1. Stimuli used in Experiments 1 & 4

The 60 trait adjectives used here were a sub-selection of the 80 trait words used in previous studies (Korn et al., 2012, 2014). Half of the selected words were positive and half were negative (according to ratings in the earlier study; Korn et al., 2012). The trait adjectives were hand-annotated according to the Big-Five categories. These hand-annotations were compared to a larger list of German trait adjectives, which relied on a slightly different factorization than the Big-Five (Ostendorf, 1990). We tried to use English translations consisting of one word. As is often the case for translations, a mixture of words could give a better idea of the word meaning (especially in the case of words with rather similar meaning).

Table from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

Item number	German trait word used	Approximate English translation
Factor 4: agreeableness – positive traits		
1	altruistisch	altruistic
2	anständig	decent
3	aufrichtig	honest
4	begnügsam	unpretentious
5	diplomatisch	diplomatic
6	ehrlich	candid
7	fair	fair
8	freigiebig	bountiful
9	generös	open-handed
10	gerecht	just
11	großmütig	magnanimous
12	großzügig	generous
13	hilfsbereit	helpful
14	kollegial	like a good colleague
15	konstruktiv	constructive
16	kooperativ	cooperative
17	loyal	loyal
18	rücksichtsvoll	considerate
19	selbstlos	selfless
20	solidarisch	solidary
21	spendabel	generous
22	treu	faithful
23	unbestechlich	incorruptible
24	uneigennützig	disinterested
25	unterstützend	supportive
26	verlässlich	dependable
27	verständnisvoll	understanding

28	vertrauenswürdig	trustworthy
29	verzeihend	forgiving
30	zuverlässig	reliable
Factor 5: conscientiousness – positive traits		
31	akkurat	accurate
32	anpackend	energetic
33	arbeitsam	industrious
34	beharrlich	persistent
35	diszipliniert	organized
36	effizient	efficient
37	eifrig	keen
38	engagiert	engaged
39	entschlusskräftig	decisive
40	fleißig	hard-working
41	flink	swift
42	fokussiert	focused
43	geduldig	patient
44	genau	meticulous
45	gewissenhaft	conscientious
46	gründlich	thorough
47	konsequent	consistent
48	leistungsorientiert	achievement-oriented
49	ordentlich	tidy
50	ordnungsliebend	orderly
51	pflichtbewusst	dutiful
52	planvoll	tactical
53	pünktlich	punctual
54	sorgfältig	diligent
55	systematisch	systematic

56	tatkräftig	dynamic
57	tüchtig	strenuous
58	verantwortungsbewusst	responsible
59	verantwortungsvoll	responsible
60	zielstrebig	goal-oriented

Table 9.2 Stimuli used in experiment 2 & 3.

Fifteen of the 60 trait adjectives in this list were selected from the list of 80 trait words used in previous studies (Korn et al., 2012, 2014). All selected words were positive. The trait adjectives were hand-annotated according to the Big-Five categories. We tried to use English translations consisting of one word. As is often the case for translations, a mixture of words could give a better idea of the word meaning (especially in the case of words with rather similar meaning).

Table from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

Item number	German sentence used	English translation
Factor 1: neuroticism / emotional stability		
1	Ich gerate leicht in Stress.	I get stressed out easily.
2	Ich bin die meiste Zeit entspannt.	I am relaxed most of the time.
3	Ich grüble über Dinge.	I worry about things.
4	Ich fühle mich selten deprimiert (bedrückt).	I seldom feel blue.
5	Ich bin leicht zu beunruhigen.	I am easily disturbed.
6	Ich bin leicht aus der Fassung zu bringen.	I get upset easily.
7	Ich wechsle oft meine Stimmung.	I change my mood a lot.
8	Ich habe häufig Stimmungsschwankungen.	I have frequent mood swings.
9	Ich bin leicht zu reizen.	I get irritated easily.
10	Ich fühle mich oft deprimiert (bedrückt).	I often feel blue.
Factor 2: extraversion		
11	Ich bringe eine Party in Schwung.	I am the life of the party.
12	Ich spreche nicht viel.	I don't talk a lot.
13	Ich fühle mich wohl unter Menschen.	I feel comfortable around people.
14	Ich halte mich im Hintergrund.	I keep in the background.
15	Ich beginne Gespräche.	I start conversations.
16	Ich habe wenig zu sagen.	I have little to say.
17	Ich spreche mit vielen verschiedenen Leuten auf Partys.	I talk to a lot of different people at parties.
18	Ich mag es nicht, Aufmerksamkeit auf mich zu ziehen.	I don't like to draw attention to myself.
19	Ich habe kein Problem im Zentrum der Aufmerksamkeit zu stehen.	I don't mind being the center of attention.
20	Ich bin still unter Fremden.	I am quiet around strangers.
Factor 3: openness to experience		
21	Ich besitze einen großen Wortschatz.	I have a rich vocabulary.
22	Ich habe Schwierigkeiten abstrakte Ideen zu verstehen.	I have difficulty understanding abstract ideas.
23	Ich habe eine lebhafte Vorstellungskraft.	I have a vivid imagination.
24	Ich bin an abstrakten Ideen nicht interessiert.	I am not interested in abstract ideas.

25	Ich habe ausgezeichnete Ideen.	I have excellent ideas.
26	Ich habe kein gutes Vorstellungsvermögen.	I do not have a good imagination.
27	Ich verstehe Dinge schnell.	I am quick to understand things.
28	Ich gebrauche schwierige Wörter.	I use difficult words.
29	Ich verbringe Zeit damit über Dinge nachzudenken.	I spend time reflecting on things.
30	Ich bin voller Ideen.	I am full of ideas.
Factor 4: agreeableness		
31	Ich empfinde wenig für andere.	I feel little concern for others.
32	Ich bin interessiert an anderen Menschen.	I am interested in people.
33	Ich beleidige andere.	I insult people.
34	Ich fühle mit anderen Menschen mit.	I sympathize with others' feelings.
35	Ich bin nicht interessiert an den Problemen anderer Leute.	I am not interested in other people's problems.
36	Ich habe ein weiches Herz.	I have a soft heart.
37	Ich bin nicht wirklich interessiert an anderen.	I am not really interested in others.
38	Ich nehme mir Zeit für andere.	I take time out for others.
39	Ich fühle die Gefühle anderer.	I feel others' emotions.
40	Ich bringe Leute dazu, sich wohl zu fühlen.	I make people feel at ease.
Factor 5: conscientiousness		
41	Ich bin immer vorbereitet.	I am always prepared.
42	Ich lasse meine Sachen herumliegen.	I leave my belongings around.
43	Ich achte auf Details.	I pay attention to details.
44	Ich verursache großes Durcheinander.	I make a mess of things.
45	Ich erledige lästige Routinearbeiten unmittelbar.	I get chores done right away.
46	Ich vergesse oft, Dinge an ihren Platz zurückzulegen.	I often forget to put things back in their proper place.
47	Ich mag Ordnung.	I like order.
48	Ich drücke mich vor meinen Pflichten.	I shirk my duties.
49	Ich folge einem Plan.	I follow a schedule.
50	Ich bin anspruchsvoll in meiner Arbeit.	I am exacting in my work.

Table 9.3. Stimuli used in experiment 5.

We used 50 items from the German translation of the IPIP (International Personality Item Pool, which consists of lexical Big-Five factor markers). In the task, sentences were changed to the third person singular (e.g., “she gets stressed out easily”).

Table from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

Experiment	n Tested	Age (mean)	Age (SD)	n Female	n Final	n profiles removed	n sessions removed
1 (Original)	35	24.37	3.66	23	35	0	1
2 (Constructed profiles)	42	23.67	3.19	28	42	0	0
3 (Two Factors)	59	25.37	4.95	30	59	0	2
4 (Fashion models)	30	25.38	4.91	15	29	1	0
5 (IPIP items)	30	25.38	4.91	15	28	2	8

Table 9.4. Overview of the participants for all five experiments

The same group of participants was used for Experiment 4 and 5.

Table from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

Experiment	Stimuli	n Factors	Profiles	n Profiles	Scale	n Trials
1 (Original)	60 traits	5	Real (online)	4	1-8	240
2 (Constructed profiles)	60 traits	2	Constructed	4	1-8	240
3 (Two Factors)	60 traits	2	Real (online)	4	1-8	240
4 (Fashion models)	60 traits	5	Real (fashion)	4	1-8	240
5 (IPIP items)	50 IPIP items	5	Real (IPIP)	5	1-5	250

Table 9.5. Overview of the stimuli

In Experiments 1–4, we used personality adjectives (i.e., trait words such as generous, diligent; see Supplementary Table 9.1 for Experiments 1 & 4 and Supplementary Table 9.2 for Experiments 2 & 3). Sixty adjectives were presented per profile (i.e., for each of the four people about which participants learned). We used traits from all five factors of the Big-Five or traits from only two factors (i.e., agreeableness and conscientiousness). In Experiment 5, for each of the five profiles, participants saw 50 items from the German translation of the IPIP (International Personality Item Pool, which consists of lexical Big-Five factor markers; see Supplementary Table 9.3). Profiles for Experiments 1 were selected from self-ratings of people from an unrelated sample of a previous lab study (Korn et al., 2012). Profiles for Experiment 2 were constructed by specifying the mean for the two factors and randomly adding noise according to a specified SD. Profiles for Experiment 3 were selected from the self-ratings of participants in Experiment 2. Profiles for Experiment 4 were selected from self-ratings given by a group of female fashion models for a related online study. All four selected persons have worked internationally as fashion models for several years. Profiles for Experiment 5 were selected from self-ratings of a large online dataset on the IPIP with over 1 million participants (Open Source Psychometrics Project; https://openpsychometrics.org/). We selected five profiles with average ratings on 4 out of the 5 factors (mean within 1 SD) but divergent scores on the remaining factor (mean above 1 SD). That is, each profile was divergent on another factor.

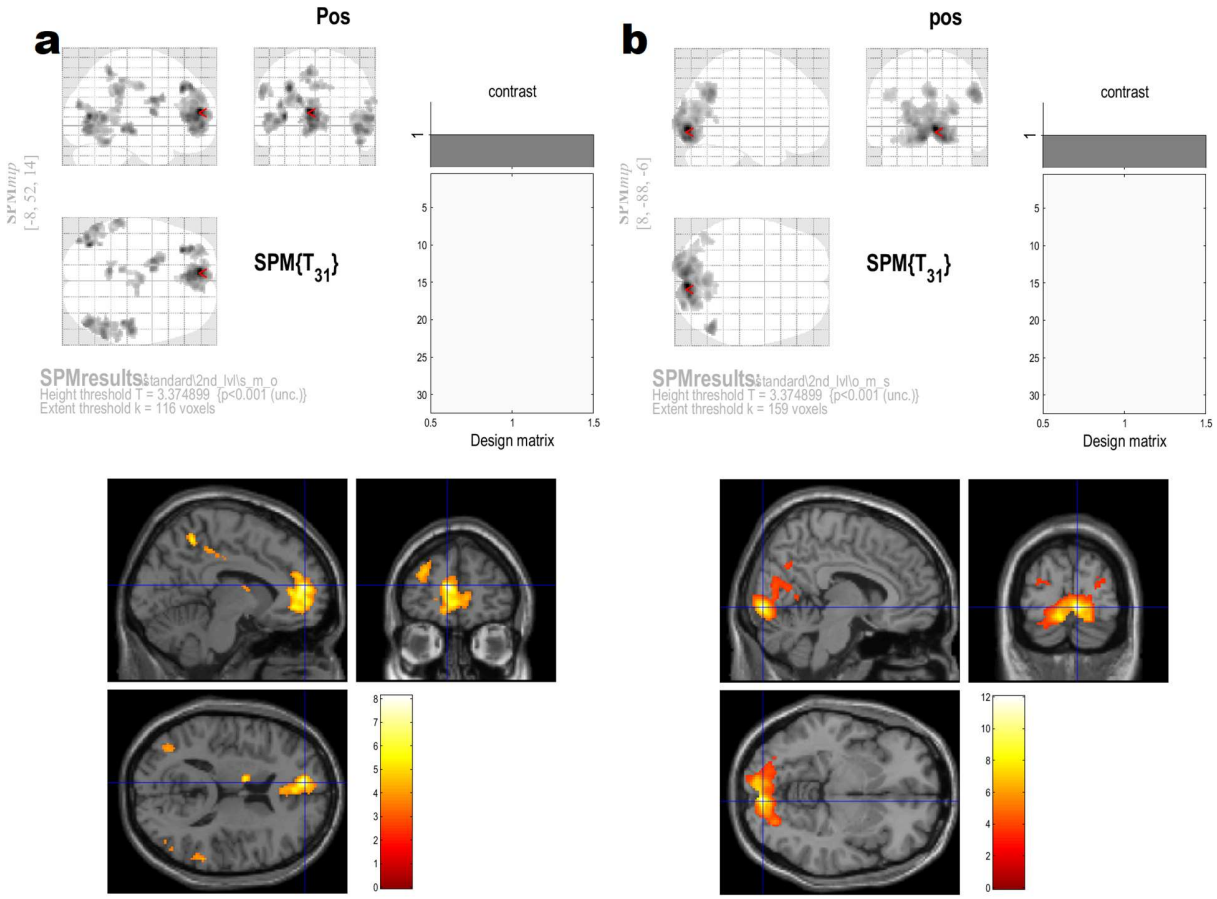
Table from: Frolichs, K.M.M., Rosenblau, G. & Korn, C.W. Incorporating social knowledge structures into computational models. Nat Commun 13, 6205 (2022), Springer Nature.

Study 2

Supplementary analysis Study 2 – Personality learning with no social interactions

Univariate

Because study 2 was a completely new dataset we performed some standard univariate analysis. In the first GLM we found a large cluster of activity in the vmPFC for the contrast self>other (Supplementary Figure 2.1A). Other clusters of activity are summarized in table 2. The opposite contrast, other>self, revealed significant clusters in the bilateral visual cortex (Supplementary Figure 1B and Supplementary table 1). The second GLM, which tested the effects of the trait word length using a parametric modulator revealed two clusters in the visual cortex as hypothesized.



Supplementary Figure 2.1. Main results from the univariate analysis

a) The contrast self > other reveals a big cluster of activity in the mPFC, **b**) The contrast other>self has a cluster of activity in the visual cortex.

	Trait word German	Trait word English	Big-5 factor
1	'aufrichtig'	Honest	Conscientiousness
2	'bescheiden'	Modest	Conscientiousness
3	'diszipliniert'	Organized	Conscientiousness
4	'effizient'	Efficient	Conscientiousness
5	'einfühlsam'	Empathetic	Agreeableness
6	'enthusiastisch'	Enthusiastic	Extraversion
7	'fleißig'	Hard-working	Conscientiousness
8	Big	Friendly	Agreeableness
9	'geistesgegenwärtig'	Quick-witted	Extraversion
10	'gelassen'	Composed	Neuroticism
11	'geschickt'	Skilled	Conscientiousness
12	'gesellig'	Sociable	Agreeableness
13	'großzügig'	Generous	Agreeableness
14	'hilfsbereit'	Helpful	Agreeableness
15	'höflich'	Polite	Agreeableness
16	'kompetent'	Competent	Conscientiousness
17	'kooperativ'	Cooperative	Agreeableness
18	'kreativ'	Creative	Openness
19	'lebenslustig'	Fun-loving	Extraversion
20	'locker'	Easy-going	Neuroticism
21	'loyal'	Loyal	Agreeableness
22	'offen'	Open-minded	Openness
23	'ordentlich'	Tidy	Conscientiousness
24	'respektvoll'	Respectful	Agreeableness
25	'scharfsinnig'	Astute	Openness
26	'schlagfertig'	Articulate	Extraversion
27	'selbstständig'	Self-reliant	Neuroticism

28	'sorgfältig'	Diligent	Agreeableness
29	'souverän'	Confident	Neuroticism
30	'spontan'	Spontaneous	Openness
31	'tatkräftig'	Dynamic	Extraversion
32	'tolerant'	Tolerant	Openness
33	'vernünftig'	Level-headed	Agreeableness
34	'verständnisvoll'	Understanding	Agreeableness
35	'vertrauenswürdig'	Trustworthy	Agreeableness
36	'vielseitig'	Versatile	Openness
37	'weitsichtig'	Perspicacious	Openness
38	'wissbegierig'	Inquisitive	Openness
39	'zielstrebig'	Goal-oriented	Conscientiousness
40	'zuverlässig'	Reliable	Agreeableness
41	'aggressiv'	Aggressive	Agreeableness
42	'ängstlich'	Anxious	Neuroticism
43	'arrogant'	Arrogant	Agreeableness
44	'bieder'	Overly conservative	Openness
45	'chaotisch'	Chaotic	Conscientiousness
46	'egoistisch'	Selfish	Agreeableness
47	'eitel'	Conceited	Agreeableness
48	'engstirnig'	Narrow-minded	Openness
49	'feige'	Cowardly	Neuroticism
50	'gehässig'	Spiteful	Agreeableness
51	'großmäulig'	Loud-mouthed	Agreeableness
52	'heuchlerisch'	Two-faced	Agreeableness
53	'hinterhältig'	Conniving	Agreeableness
54	'humorlos'	Humorless	Extraversion
55	'inkonsequent'	Inconsistent	Conscientiousness

56	'kalt'	Cold-hearted	Extraversion
57	'launisch'	Moody	Neuroticism
58	'leichtsinnig'	Foolhardy	Conscientiousness
59	'nachtragend'	Unforgiving	Agreeableness
60	'naiv'	Naive	Agreeableness
61	'oberflächlich'	Superficial	Conscientiousness
62	'opportunistisch'	Opportunistic	Agreeableness
63	'pedantisch'	Pedantic	Conscientiousness
64	'rücksichtslos'	Inconsiderate	Agreeableness
65	'scheu'	Unassertive	2
66	'stur'	Stubborn	Agreeableness
67	'träge'	Lazy	Openness
68	'unentschlossen'	Indecisive	Neuroticism
69	'ungeduldig'	Impatient	Agreeableness
70	'unnahbar'	Inapproachable	Extraversion
71	'unpünktlich'	Tardy	Conscientiousness
72	'unsicher'	Insecure	Neuroticism
73	'unsympathisch'	Unpleasant	Agreeableness
74	'verschwenderisch'	Wasteful	Conscientiousness
75	'voreilig'	Rash	Conscientiousness
76	'voreingenommen'	Biased	Openness
77	'wehleidig'	Whiny	Neuroticism
78	'zickig'	Catty	Agreeableness
79	'zwanghaft'	Obsessive	Conscientiousness
80	'zynisch'	Cynical	Neuroticism

Table 9.6 Trait words used in the task for the self- and other-ratings for study 1.

	Trait word German	Trait word English	Big-5 factor	Pos/ Neg
1	aufrichtig	sincere	Conscientiousness	Positive
2	bescheiden	humble	Conscientiousness	Positive
3	enthusiastisch	enthusiastic	Extraversion	Positive
4	fleißig	hard-working	Conscientiousness	Positive
5	freundlich	friendly	Agreeableness	Positive
6	gelassen	serene	Neuroticism	Positive
7	großzügig	generous	Agreeableness	Positive
8	hilfsbereit	helpful	Agreeableness	Positive
9	höflich	polite	Agreeableness	Positive
10	kreativ	creative	Openness	Positive
11	lebenslustig	fun-loving	Extraversion	Positive
12	locker	easy-going	Neuroticism	Positive
13	ordentlich	orderly	Conscientiousness	Positive
14	schlagfertig	quick-witted	Extraversion	Positive
15	selbstständig	independent	Neuroticism	Positive
16	souverän	sovereign	Neuroticism	Positive
17	spontan	spontaneous	Openness	Positive
18	tatkräftig	energetic	Extraversion	Positive
19	tolerant	tolerant	Openness	Positive
20	vielseitig	versatile	Openness	Positive
21	ängstlich	timid	Neuroticism	Negative
22	arrogant	arrogant	Agreeableness	Negative
23	egoistisch	egotistical	Agreeableness	Negative
24	engstirnig	narrow-minded	Openness	Negative
25	hinterhältig	conniving	Agreeableness	Negative
26	humorlos	humorless	Extraversion	Negative
27	kaltherzig	cold-hearted	Agreeableness	Negative

28	launisch	capricious	Neuroticism	Negative
29	leichtsinnig	reckless	Conscientiousness	Negative
30	pedantisch	pedantic	Conscientiousness	Negative
31	scheu	timid	Extraversion	Negative
32	stur	stubborn	Agreeableness	Negative
33	träge	sluggish	Openness	Negative
34	unentschlossen	indecisive	Neuroticism	Negative
35	unnahbar	unapproachable	Extraversion	Negative
36	unpünktlich	unpunctual	Conscientiousness	Negative
37	unsympathisch	unsympathetic	Agreeableness	Negative
38	voreingenommen	prejudiced	Openness	Negative
39	wehleidig	sniveling	Neuroticism	Negative
40	zwanghaft	compulsive	Conscientiousness	Negative

Table 9.7 Trait words used in the task for the self- and other-ratings for study 2.

Study 3

General overview of the tasks

Gitter consisted of 9 tasks of which 1 & 2 were only performed on day one, 3-8 were repeated on both days and the 9th task was only performed on day two. Moreover, on day two, task 7 was changed slightly and performed in the scanner. The tasks 1-4 and the behavioral part of task 7 had the same set-up, which will be described briefly (Figure 1). On the right side of the screen are the sliders, these two horizontal bars (colored blue and red) can be controlled in a vertical direction by using the keys 'k' and 'm' for up and down of the right (red) slider and 'a' and 'z' for up and down for the left (blue) slider. The sliders always initialize in the center and can move until the bigger circles at the top and bottom with the smaller circles functioning as rough guidelines. The settings of the sliders will indicate both movement direction and velocity of the bars, positioned on the left half of the screen. The sliders at the center indicate no movement, above the center indicates the bars moving up and below the center indicates downward movement. The distance away from the center indicates velocity of movement. Participants are free to move the sliders and can confirm their choice using the 'spacebar' which will cause the bars to start moving, bars always move for 2 seconds total. The black horizontal lines indicate the minimum and maximum of the bars' allowed movement. Throughout the experiment (except for task 1) the height of the bars will reflect scores on two personality traits (generous and diligent).

During the behavioral training, our aim is twofold. First, we assess participants' training progress on tasks that are similar to the task performed in the scanner. We want to make sure that participants are proficient in navigating the trait space, i.e., in indicating whether and how the bar movements relate to the traits of six strangers. Training progress on most tasks is reflected in participants' choice accuracy, reaction time, and number of tries during the training tasks. Second, we want to ensure that participants actually learn about and consider the strangers' personality traits (and not just the bar movements independent of their reflection of a personality trait). Therefore, there are a number of tasks that evaluate participants' view of the strangers as a whole (e.g., as more than two-dimensional personalities) by comparing them to their acquaintances, rating them on different personality traits from the same factors and explicitly placing them on a map.

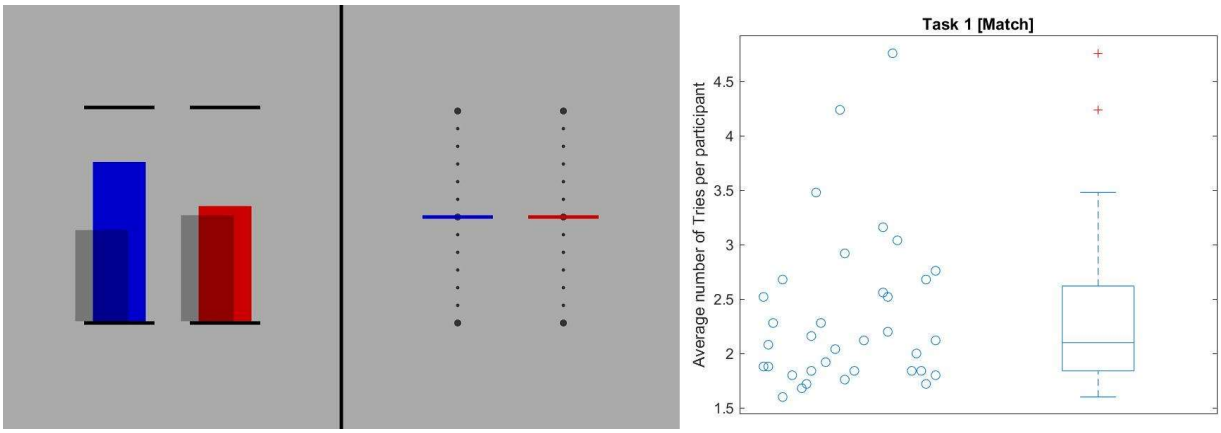
With the exception of task 9, participants were never made aware that the task could be represented on a 2-dimensional plane. Furthermore, instructions never used words pertaining to navigation or something similar.

Arena and stranger placing

The "arena" in which the strangers were hidden consisted of a 100x100-pixel grid. That is, there are 100 equal steps between the minimum and maximum on both axes. Changing the height of the bars meant one was traversing through the arena. The strangers were placed in the arena according to a semi-random process.

Task 1 – Match

In task 1 [match], the goal was to get participants used to the controls. Participants were asked to match the length of the colored bars to their respective shadowed counterpart that was slightly overlapping on the left. Moving the bars happened in two steps: first, the sliders would be moved to a certain position that would indicate the direction and velocity of the bars. Second, the settings would be confirmed by pressing the spacebar which would cause the bars to start moving in the given direction. Participants got unlimited trials to match the opaque bars, bars didn't reset between tries. A trial was completed successfully when participants got both colored bars to match the opaque bars within a 5% error margin. The match task required a total of 25 successful trials.



Supplementary Figure 3.1. Overview of the match task and participants' results

Left, to get used to the controls participants had to match the colored bars to the transparent gray bars superimposed on them. Right, the average number of tries (moving the bars) for each participant to match the colored bars to the transparent gray bars.

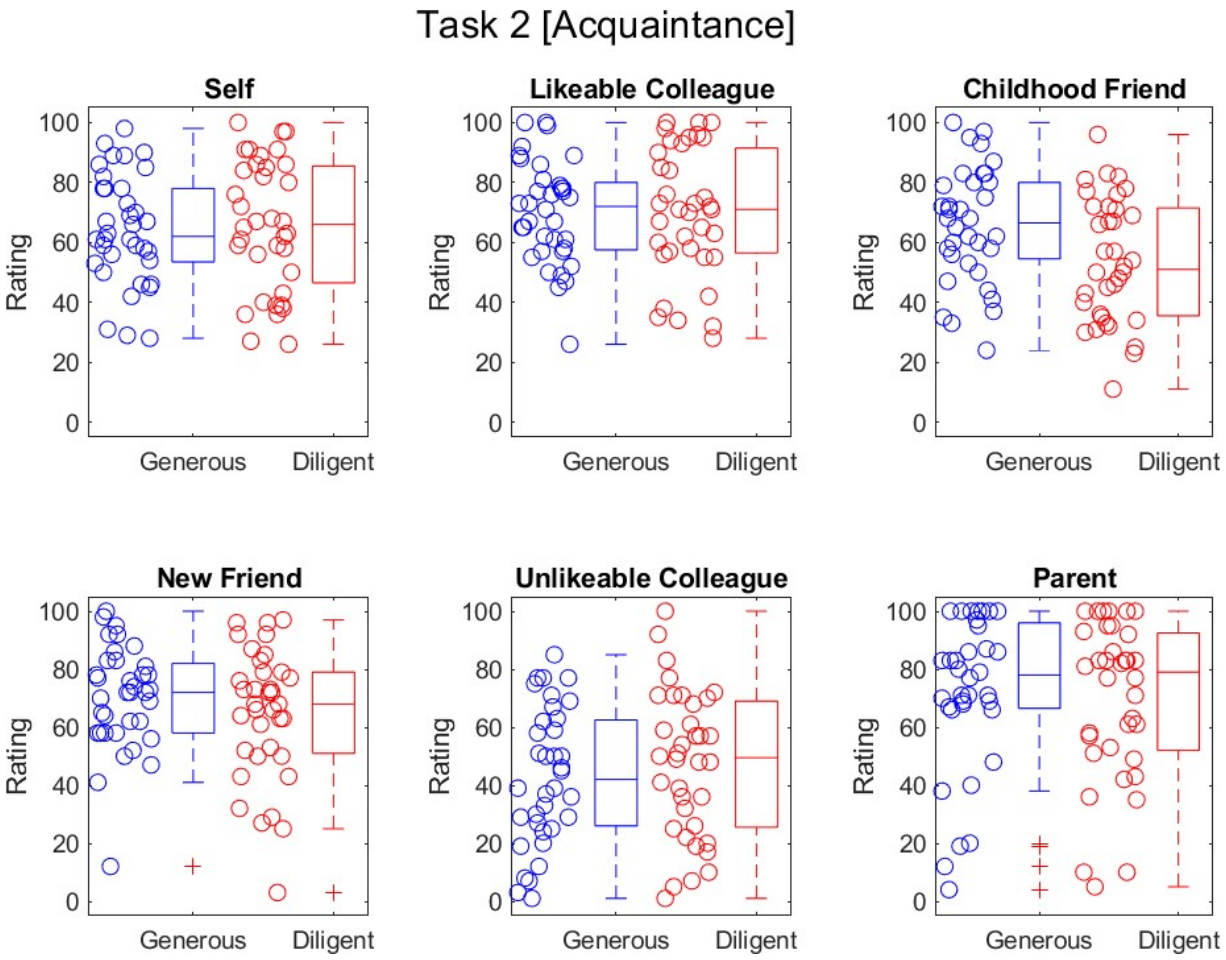
Task 2 – Acquaintances

After participants learned to manipulate the bars' positions, task 2 [acquaintances], introduced a meaning to the bars' height. Both bars indicated a personality trait (i.e., generous for the blue bar and diligent for the red bar), the height of the bar reflects a person's rating on the personality traits, where a full-length bar reflects a high score on the trait (i.e., fully possessing the trait) and a low bar indicated a low score (i.e., not possessing the trait). This coupling of the traits generous and diligent to their respective bars was applied throughout the whole experiment.

To get participants familiar with this meaning of the bars we asked them to rate six acquaintances using the bars. Acquaintances roles were given to participants to help them with their selection as well as trying to get acquaintances that would be rated differently (e.g., one would expect friends to be rated very similar) (Supplementary Table 1).

The task proceeded as follows for all six acquaintance roles: 1) Participants received written instructions where the meaning of the bars was explained. 2) Participants were presented with one of six acquaintance roles and had to imagine an acquaintance that fit this role. 3) They could enter this acquaintance's first name using the keyboard. 4) Participants were given control of the sliders and were asked to set the bars to the position that reflected this specific acquaintance's score on generous and diligence. They were free to move the bars for as long as they needed. 5) When satisfied participants could indicate through button press after which a next trial started or, when all six acquaintances were rated, the task ended.

We made sure participants rated these acquaintances different from the medium i.e., both bars at half height.



Supplementary Figure 3.2. Participant acquaintance ratings.

Each circle represents a participant's rating on the two traits generous (blue) and diligent (red) for six acquaintances. Each acquaintance is in a separate plot.

Task 3 – Explore

After successful completion of tasks 1 and 2, we expected participants to be comfortable with the controls as well as with the bars reflecting the personality traits. For task 3 [explore] the participants were tasked with finding and remembering 6 unfamiliar people (hereafter called strangers) based on their personality ratings on the two traits generous and diligent (i.e., bar height). The strangers were hidden, in that participants were initially unaware of their ratings. Participants could find the strangers by maneuvering the bars in the correct position i.e., in the same position as a strangers' rating (within 6% error margin in all directions). When a stranger was found their name would pop-up in the middle of the screen. To keep track, the six names were displayed in the top left of the screen and a green checkmark behind the name indicated the participant had found this stranger

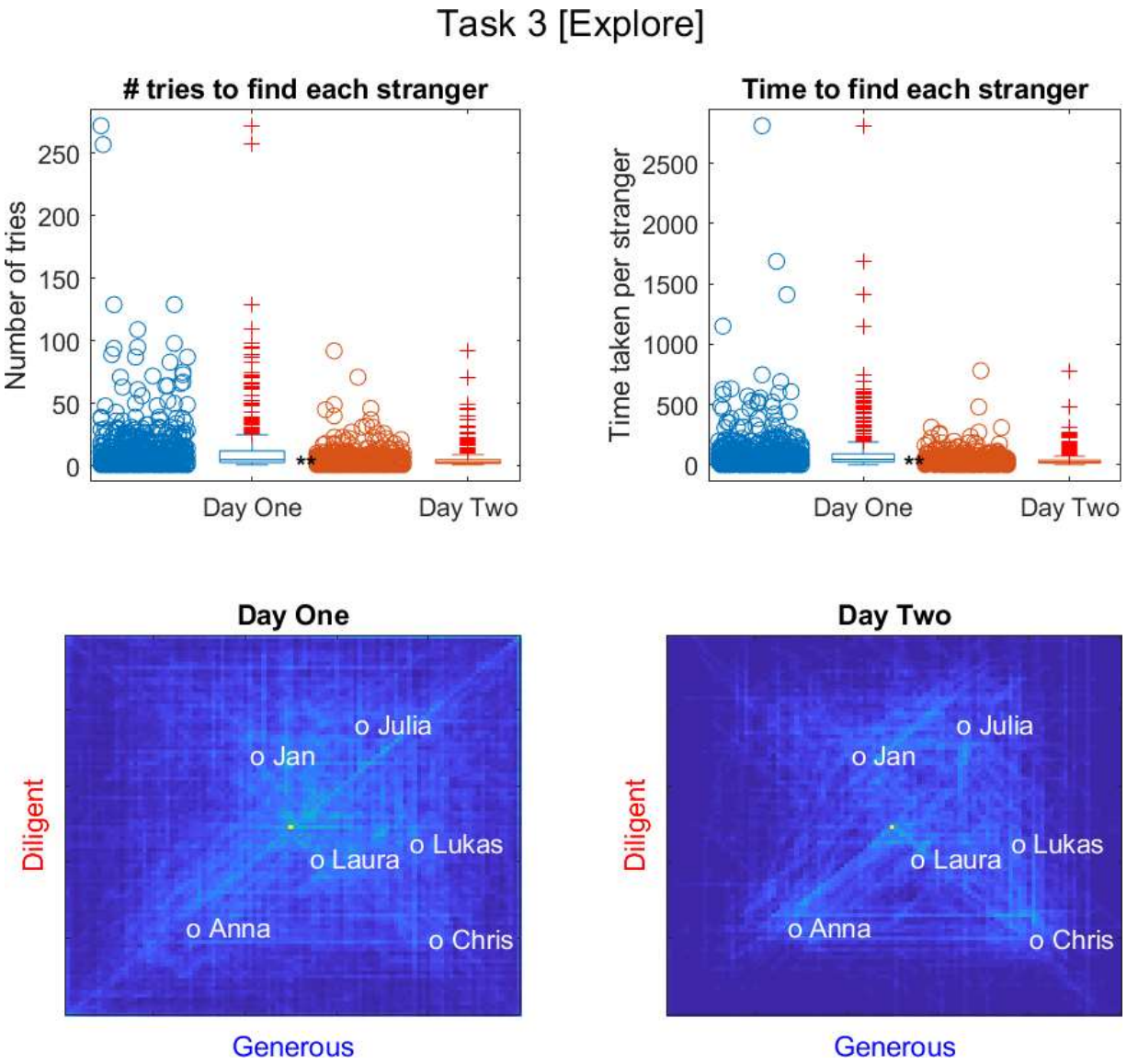
once. Participants were instructed to find all strangers 3 times on separate occasions i.e., finding at least one other stranger in between.

Performance improvements from day one to two

Participants performed task 3 on both days, we expected participants to improve their performance significantly from day one to two. Improvement was measured both on the time taken per trial as well as number of tries (i.e., number of times participants had to move the bars until they successfully found a stranger). We expected that the number of tries would be significantly less from day one to two, and thus computed a right-tailed two-sample t-test for day one (M= 11.94, SD= 21.64) vs day two (M= 4.75, SD= 7.11) and found that the number of tries indeed was significantly smaller $t(1286) = 8.00$, $p < .001$ on day two. Furthermore, we computed a right-tailed two-sample t-test for time (seconds) taken per trial on day one (M= 87.17, SD= 169.21) and two (M= 35.72, SD= 51.08) and found that participants needed significantly less time on day two $t(1280) = 7.37$, $p < .001$. These results show that participants significantly improved their performance from day one to two.

Participants' movement through personality space

Evaluation of the angles made when moving through the personality space revealed an interesting pattern. By far the most movements participants made were in one of four cardinal directions i.e., only one bar moving, this was followed by movements made in the four ordinal directions i.e., both bars moving with the same velocity. Overall participants' movements got more focused around the 6 names in the trait space from day one to two. The borders, where no strangers were positioned, were visited a lot less.



Supplementary Figure 3.3. Overview for task 03.

Top, both the number of tries (left) as well as the time taken (right) on each trial (finding a stranger) decreased from day one to day two. Bottom, participants' movement in the trait space concentrated more around the strangers on day two.

Task 4 – Collect

Task 4 [collect] assumed participants roughly learned the strangers' generosity and diligence in the previous task. Therefore, they were now asked to navigate to specific people from randomized start positions (i.e., differing bar heights). At the start of a trial one of the strangers' names was presented in the center of the screen, bar heights were semi-random, in that the heights were randomized but adjusted so that each name was reachable in one trial (movement of the bars). Like task three, a

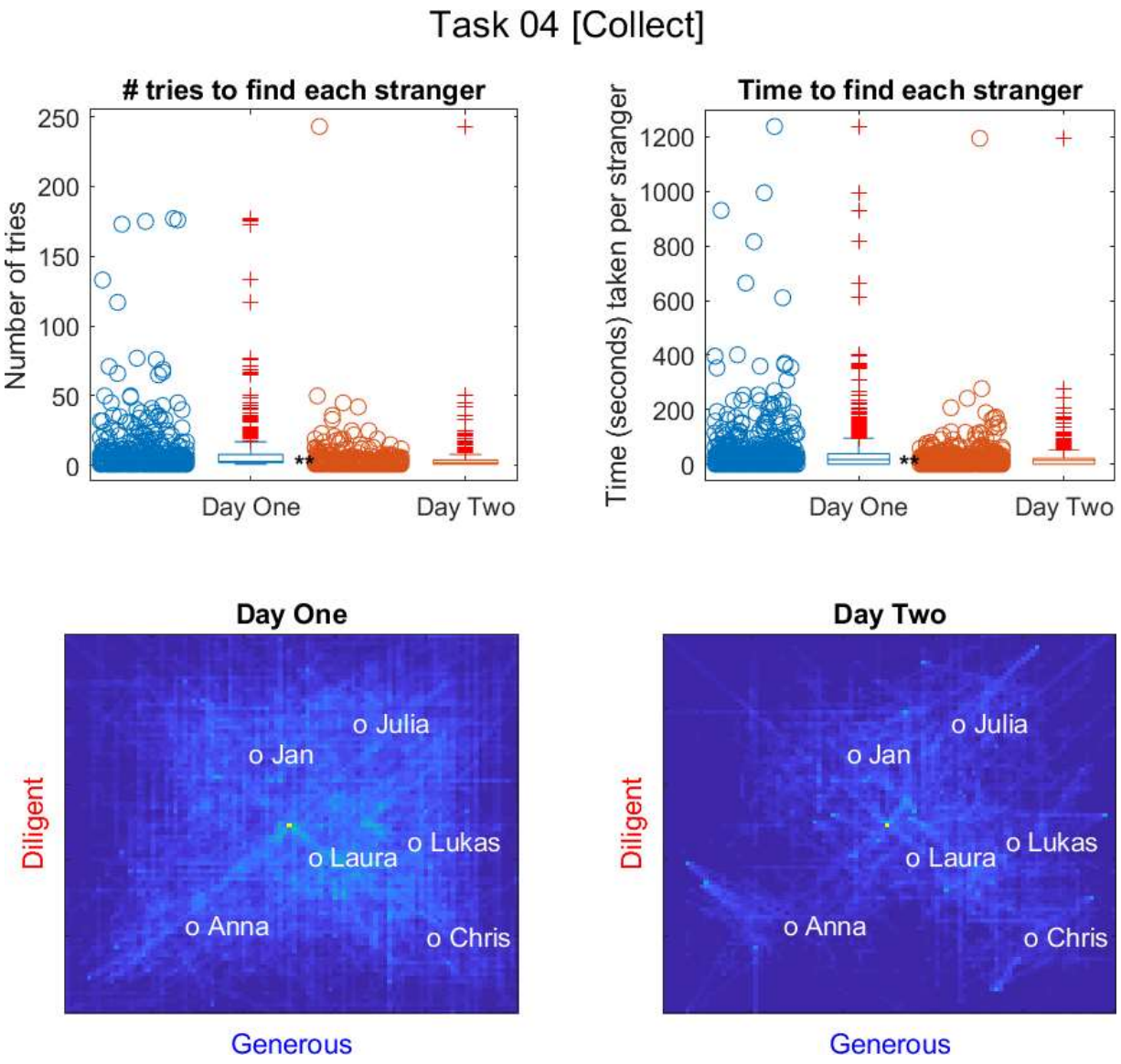
name was counted as ‘collected’ when participants reached the right bar heights (trait ratings) with a 6% error margin.

Performance improvements from day one to two

To check whether participants’ performance increased from day one to two, we computed a right-tailed two-sample t-test for both time per trial and total amount of tries per trial. Comparing time per trial (seconds) on day one (M= 63.39, SD= 126.73) vs two (M= 31.88, SD= 71.22) we found a significant decrease $t(686) = 4.02$, $p < .001$. Furthermore, number of tries (i.e., bar height adjustments) per trial on day one (M= 11.34, SD= 24.77) vs two (M= 4.81, SD= 12.11) also showed a significant decrease $t(1287) = 6.01$, $p < .001$. Indicating that participants’ performance increased from day one to two and participants most likely remembered the strangers’ trait ratings more accurately on day two.

Participants’ movement through personality space

Like task 3, participants’ movements got more focused around the actual positions of the strangers.



Supplementary Figure 3.4. Overview for task 04.

Top, both the number of tries (left) as well as the time taken (right) on each trial (finding a stranger) decreased from day one to day two. Bottom, participants' movement in the trait space concentrated more around the strangers on day two.

Task 5 – Simple Similarities

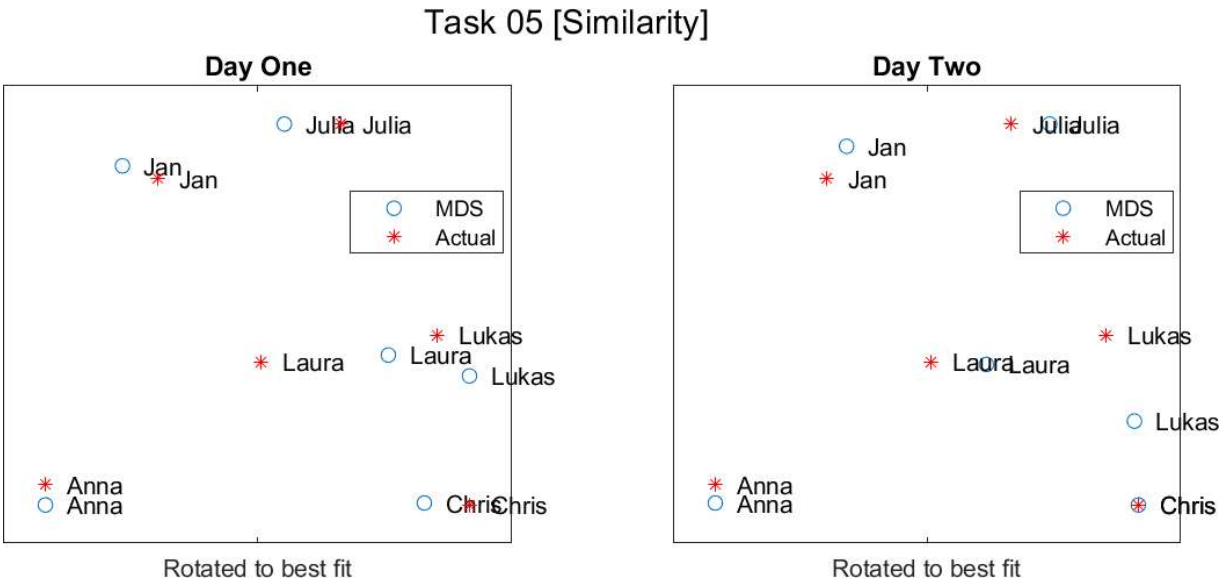
With the previous tasks done, we expected the participants to have internalized the strangers' traits. In task 5 [simple similarities] we asked participants to rate the similarities within and between their acquaintances and the strangers. Every trial participants were presented with a pair, from a total of 12 persons (6 strangers and 6 acquaintances) every person was paired with 11 others for a total 66

trials. Every pair was presented only once. During the trial participants were asked to rate the pairs' similarity on a scale from 1 'not similar at all' to 7 'very similar'.

To get an idea of how participants rated the similarities between the strangers we used multidimensional scaling (MDS) on averaged similarity ratings per day (i.e., the similarity ratings for all six strangers averaged over all participants). The coordinates retrieved from MDS and those from the strangers were normalized to get them in the same scale. After this the retrieved MDS coordinates were rotated to their closest fit to the strangers' coordinates. On both days the MDS coordinates closely fitted the actual coordinates.

Next we quantified participants' accuracy on the similarity ratings by comparing them with the actual Euclidean positions of both the strangers and acquaintances. For both groups separately we calculated a Euclidean distance matrix between all pairs from their explicit positions (i.e., coordinates of the trait space that were determined by us for the strangers and rated by participants for the acquaintances). We then calculated the correlations between these distance matrices and the similarity ratings for these separate groups for both days. The strength of the correlation indicates how well the explicit and rated distances overlap (correlations underwent a Fisher transformation before running any statistical tests). Since participants already had a fully formed trait space of their acquaintances before the experiment we expected their accuracies to be similar on both days. This in contrast with the strangers, where we expected the (accuracy of the) similarity ratings to improve from day one to two indicating that participants got to know the participants better.

A paired-samples t-test was conducted to compare the accuracy between correlations of similarity ratings and real positions for participants' acquaintances on day one ($M = .76$, $SD = .30$) versus day two ($M = .74$, $SD = .39$). The t-test indicated no significant difference between the accuracies on day one vs day two, $t(35) = 0.60$, $p = .56$. We did expect an increase in accuracy on the second day for the strangers, therefore we calculated a left-tailed paired-samples t-test. Accuracy on day one ($M = .44$, $SD = .35$) was significantly lower than day two ($M = .76$, $SD = .31$), $t(35) = -3.99$, $p < .001$, indicating that participants got better at estimating the strangers' similarities. Moreover, accuracy for strangers on day two was comparable to that for the acquaintances on both days indicating participants resembled the trait space of the strangers in a similar manner to that of their acquaintances.



Supplementary Figure 3.5. Using MDS to represent participants' similarity ratings

Participants rated the similarities between all strangers, we used multidimensional scaling (MDS) to project these ratings on the trait space. Because this projecting can be in any orientation it was rotated until the best Euclidean fit was found with the actual position of the strangers on the trait space, this best fit is shown in the figure. Participants were accurate for both day one and two.

Task 6 – Complex Similarities

Similar to task 5, task 6 [complex similarities] tasked participants with rating similarities between people. The difference with task 5 is that similarities of pairs were rated e.g., Julia & Chris vs Anne & Lukas and ratings were only performed for strangers, with the explicit question to indicate which of the two pairs was more similar. Answers were given by the right or left arrow key. All pairs were compared once for a total of 45 trials.

Task 8 – IPIP

Task 8 [IPIP], was administered to test whether participants learned about the strangers as having more than just the two single traits they learned about in the previous tasks. On both days participants were asked to rate the strangers and acquaintances on five other traits from agreeableness and conscientiousness (i.e., the two Big-5 factors that generous and diligent fall under). The five traits used were different for the two days (see supplementary table 2). A trial consisted of the name of the person that was rated, preceded by the text 'acquaintance from another person' when it was a stranger and 'your acquaintance' when it was an acquaintance of the participant together with an item from the IPIP. Participants were tasked with rating the person on

this current item on a scale from 1 (does not apply at all) to 5 (does apply very much). Presentation of acquaintances, strangers and the order of the items were randomized for every participant.

We expected participants to have learned enough about the strangers to have their traits somewhat internalized, and thus have these current trait ratings correspond to actual ratings from both strangers and acquaintances. In order to test this we computed correlations between the 'actual' ratings on generous and diligent and the mean ratings on their corresponding factors (agreeableness and conscientiousness) separately. The correlations for the strangers were significant for both agreeableness & generous (rho .53, $p < .001$) and conscientiousness & diligent (rho .63, $p < .001$). Likewise, correlation for the acquaintances were significant for agreeableness & generous (rho .76, $p < .001$) and conscientiousness & diligent (rho .80, $p < .001$). Indicating participants saw the strangers as more than just two bars and more like humans with multiple traits.

Task 9 – Map Task

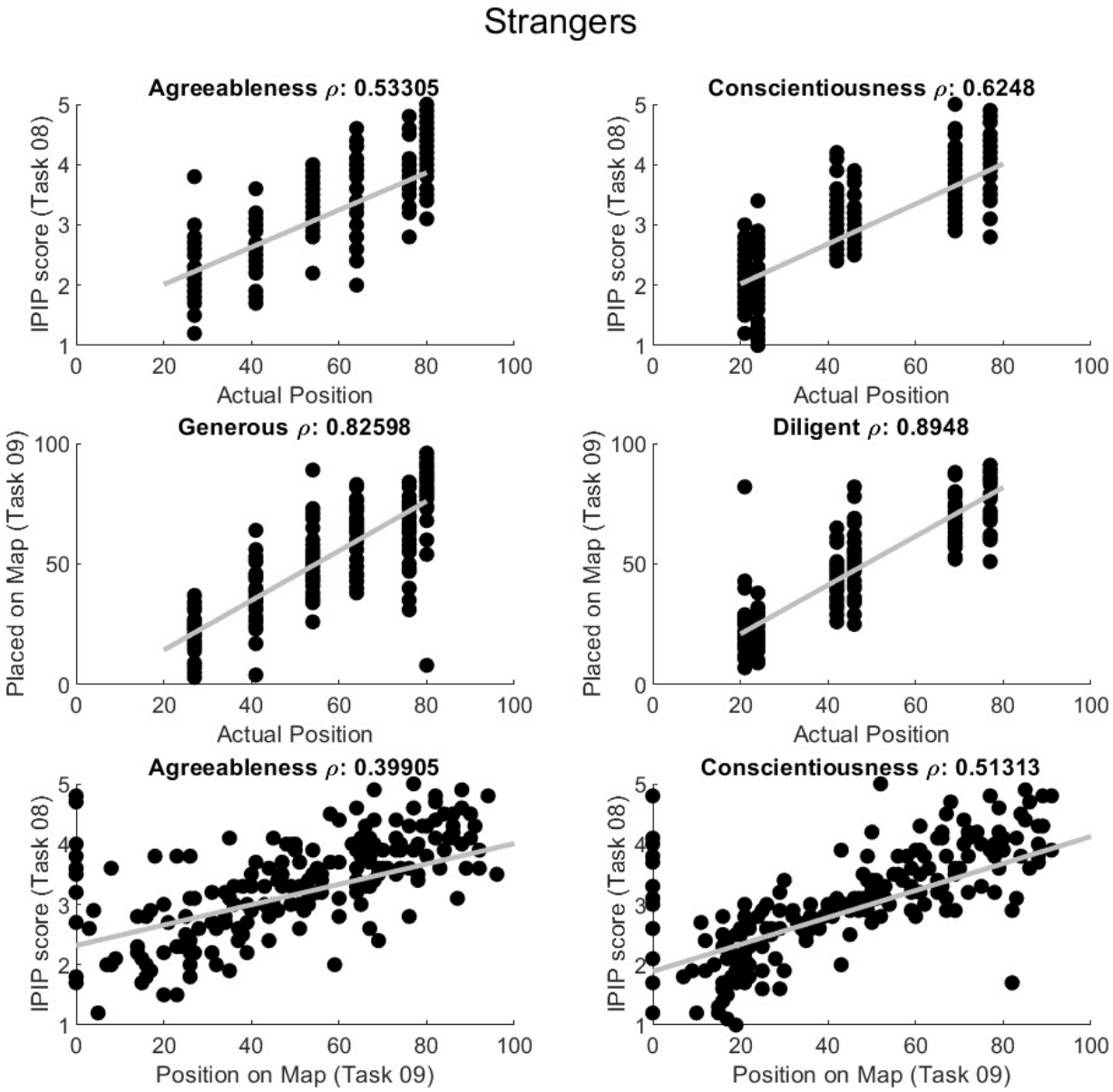
The final task, task 9 [map] was only conducted on the second day and was the last task. In this final task, it was revealed that participants could have represented the strangers on a Euclidean plane. The scores on both traits (generous and diligent) would then function as the axes. To test whether participants could use this type of representation we asked them to drag all 12 people (6 strangers and 6 acquaintances) onto their respective coordinates on the map. The map was centered on the screen with the axes clearly labeled for generous and diligent, the 12 people were ordered on either side of the map in two vertical lines of six people each. Participants used their mouse to drag every name on the map. They had unlimited time to complete the task and could click on a button labeled 'okay' to finish the task.

To test whether participants could use the map representation correctly we compared the 'actual' positions with those from the map task. For the strangers both agreeableness and & generous was significantly correlated (rho .83, $p < .001$) as well as conscientiousness & diligent (rho .90, $p < .001$). Furthermore, the positions for acquaintances were also significantly correlated for both agreeableness and & generous (rho .82, $p < .001$) and conscientiousness & diligent (rho .77, $p < .001$). Indicating that participants could use the Euclidean representation in a meaningful manner.

Finally, we compared the positions from the map task with the ratings on the IPIP factors agreeableness and conscientiousness from task 8 (Figure 11). The correlations for both the strangers and acquaintances were significant for all comparisons. Strangers: agreeableness: rho .81,

p < .001 & conscientiousness: rho .85, p < .001. Acquaintances: agreeableness: rho .83, p < .001 & conscientiousness: rho .85, p < .001.

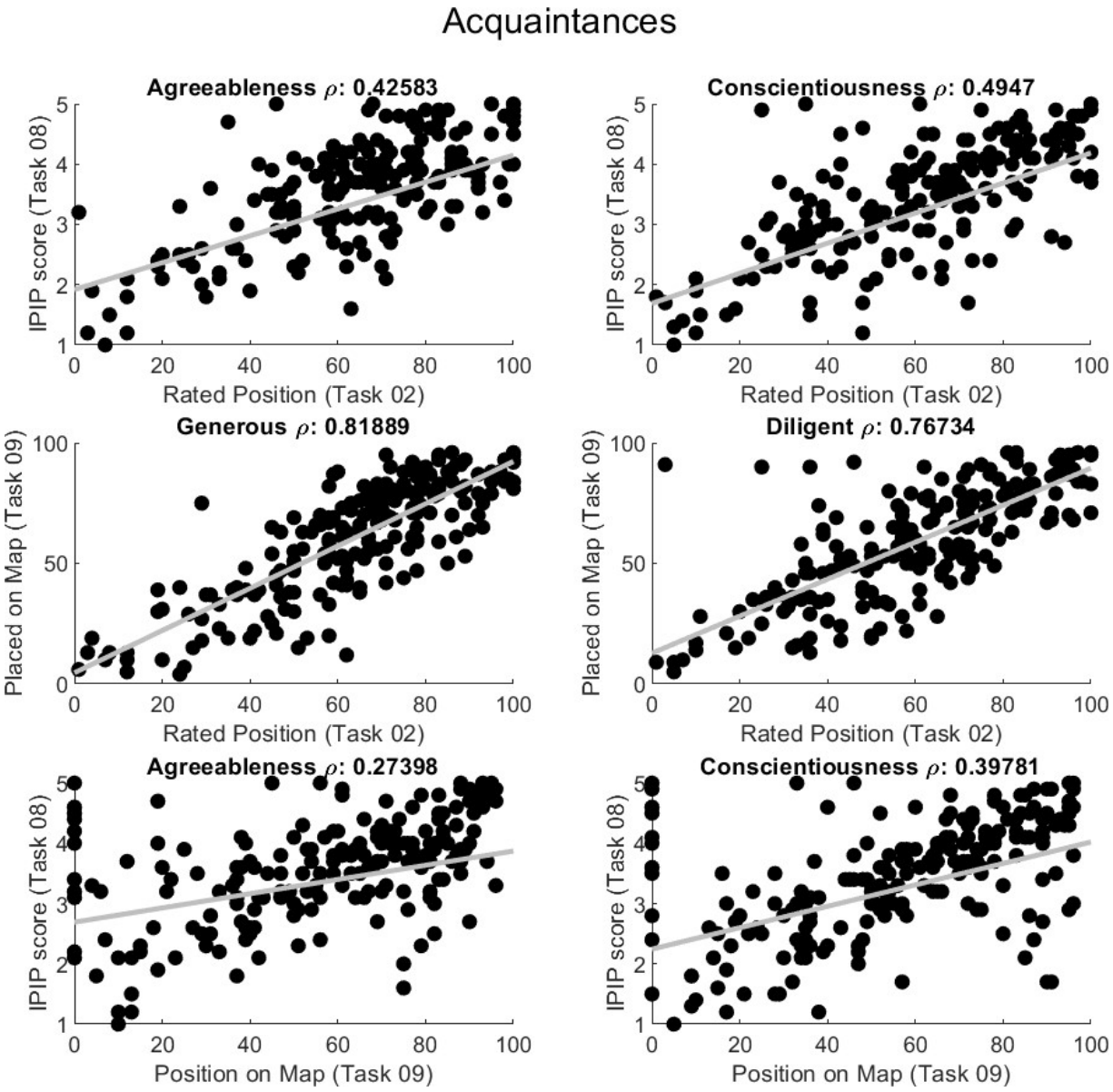
All results combined from task 5, 8 and 9 give us confidence that participants perceived the strangers as people. Especially considering that performance was often on similar levels of accuracy for both the strangers and acquaintances.



Supplementary Figure 3.6. Comparison for tasks 08 and 09 with the actual positions of the strangers.

Top, participants gave multiple ratings on trait items from the same Big-5 factor. We correlated these item ratings with the trait ratings from the strangers and found they correlated well for both agreeableness and conscientiousness. Middle, in task 09 participants were asked to drag each participant on a map with the axes generous and diligent. We

correlated the place on this task with the strangers' actual positions and found high correlations for both traits generous and diligent. Bottom, here we correlated the results from task 08 and 09, that is, the IPIP scores with the place on the map. Correlations were positive but lower than for the tasks correlated with the actual positions.



Supplementary Figure 3.7. Comparison for tasks 08 and 09 with the actual positions of the acquaintances.
Top, participants gave multiple ratings on trait items from the same Big-5 factor. We correlated these item ratings with the trait ratings the participants gave in task 02 and found they correlated well for both agreeableness and conscientiousness. Middle, in task 09 participants were asked to drag each participant on a map with the axes generous and diligent. We correlated the place on this task with participants' ratings from task 02 and found high correlations

for both traits generous and diligent. Bottom, here we correlated the results from task 08 and 09 for the acquaintances. Correlations were positive but lower than for the tasks correlated with the actual positions.

	German	English
1	sich selbst	yourself
2	einen Kollegen/Studienpartner oder eine Kollegin/Studienpartnerin, mit dem oder der Sie gerne zusammen arbeiten	a colleague or study-partner with whom you like to work together.
3	einen Freund oder eine Freundin aus der Kindheit	a friend from your childhood
4	einen aktuellen Freund oder eine aktuelle Freundin, den oder die Sie seit kurzem kennen	a current friend whom you have known since a short amount of time
5	einen Kollegen/Studienpartner oder eine Kollegin/Studienpartnerin, mit dem oder der Sie NICHT gerne zusammen arbeiten	a colleague or study-partner with whom you do not like to work together with.
6	einen Elternteil (Vater oder Mutter)	a parent (father or mother)

Table 9.8. Role of acquaintances from task 2.

Agreeableness		**Conscientiousness**	
Day One	**Day Two**	**Day One**	**Day Two**
Ich mache mir um andere Sorgen	Ich liebe es anderen zu helfen	Ich verstehe es Dinge zu bewältigen	Ich mache mehr, als von mir verlangt wird
Ich glaube an das Gute im Menschen	Für andere nehme ich mir keine Zeit	Ich arbeite hart	Ich vertrödele meine Zeit
Ich habe Mitleid mit denen, den es schlechter geht als mir	Ich bin der Meinung, dass Leute für sich selber sorgen sollten	Ich tue gerade so viel, dass es ausreicht	In meine Arbeit stecke ich nur wenig Zeit und Anstrengung
Ich misstraue anderen	Ich glaube, dass andere gute Absichten verfolgen	Ich bin stets vorbereitet	Situationen schätze ich falsch ein
Die Gefühle anderer sind mir gleichgültig	Ich zeige Mitgefühl für Obdachlose	Ich schließe Aufgaben stets erfolgreich ab	Aufgaben beginne ich stets rechtzeitig

Table 9.9. German IPIP items for task 8.

Curriculum Vitae

Job Experience ----------

2019 - 2024, Heidelberg & Hamburg **PhD Candidate**, Institute for systems neuroscience.
* PhD Topic: we focus on how humans learn about others. Through computational models we aim to uncover what potential strategies humans use. Furthermore, using functional magnetic resonance imaging, we explore how these social structures are represented in the brain.

2017 – 2018, Venlo **Software Developer**, Transceptor Technology
* Developing and maintaining software packages mainly focused on e-commerce.

2014 – 2017, Maastricht **Tutor**, Maastricht University
* Facilitate learning by supporting, guiding, and monitoring the learning process in small problem based learning groups.

2014 – 2015, Maastricht **Vice-President**, Msav Uros
* Updating and reviewing social media and website, organizing events and activities, and aiding the president (and other board members) where necessary.

Education ----------

2017-2018, Sittard **Software Developer**, EduCom Automatiserings Opleidingen.
Topic: Fast course to learn the basic skills of software development.

2015-2017, Maastricht **Research Master**, Cognitive Neuroscience, Maastricht University
Topic: Reconstructing Complex Visual Shapes from Activation in Early Visual Cortex: A Gradient-Descent Approach to Pixelwise Reconstruction

2012-2015, Maastricht **Bachelor Psychology**, Maastricht University (*Research Minor)*
Topic: Testing the suitability of function Near-Infrared Spectroscopy for a multiple choice brain-computer interface paradigm.

Teaching ----------

11.2021, Heidelberg, Germany 4EU+ (one day course on programming principles in Python)
05.2022, Heidelberg Germany 4EU+ (three day course on computational psychiatry).

Invited Talks ----------

07.2019, Salzburg, Austria Samba, Young Scientists Symposium
09.2020, Washington DC, USA Brown Bag Talk (Online)
02.2022, Germany ZIHub, mental health alliance (Online)
06.2022, Freiburg Germany Psychologie und Gehirn (PuG),

Posters ----------

07.2019, Salzburg, Austria Salzburg Mind-Brain Annual Meeting
09.2019, Berlin, Germany Conference on Cognitive Computational Neuroscience
10.2021, Mannheim, Germany ZIHub, mental health alliance

04.2022, Groningen, the Netherlands Cognitive Modelling spring school

Summer Schools ----------

07.2020, Vienna, Austria Pattern Recognition in Neuroimaging.
04.2022, Groningen, the Netherlands Cognitive Modelling

Skills

----- Organization of scientific meetings ----------

Both meetings were organized in and for the Institute of systems neuroscience in Hamburg, Germany.

2019 - 2021, Mathematics meeting. This was an informal weekly meeting with a focus on linear algebra. My job was to prepare a new topic every week and try to keep the group on track by sending reminders and summaries of our progress.

2019 - 2021, Methods meetings. This was a bi-weekly meeting focusing on a different topic of interest each week (e.g. Open Science, Pupil Data, and Software Packages). My main activities were to organize speakers, book rooms (physical & online), send invitations.

----- Data Acquisition ----------

Brain imaging	Functional and structural MRI (3 tesla), data analysis using MATLAB and SPM.
Behavioral	Programming tasks using MATLAB and Psychtoolbox
Data analysis	Advanced statistical skills

----- **Organizational** ----------

MS Office	Word, Excel, PowerPoint, Outlook
Google Drive	Docs, Sheets, Forms, Slides

----- **Other** ----------

Programming	MATLAB Python Linux (basics)
Languages	Dutch – Native English – Very Good German – Basic.
Interests	Bouldering, Mountain biking, Reading, Making (stuff).

Eidesstattliche Versicherung

Ich versichere ausdrücklich, dass ich die Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen einzeln nach Ausgabe (Auflage und Jahr des Erscheinens), Band und Seite des benutzten Werkes kenntlich gemacht habe.
Ferner versichere ich, dass ich die Dissertation bisher nicht einem Fachvertreter an einer anderen Hochschule zur Überprüfung vorgelegt oder mich anderweitig um Zulassung zur Promotion beworben habe.
Ich erkläre mich einverstanden, dass meine Dissertation vom Dekanat der Medizinischen Fakultät mit einer gängigen Software zur Erkennung von Plagiaten überprüft werden kann.

Unterschrift: