universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Theory and Application of Bayesian Inference in Social Neuroscience

verfasst von | submitted by

Mag.rer.nat. Lukas Lengersdorff BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 605 298
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Psychologie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Claus Lamm Predrag Petrovic Ass.-Prof. Mag. Dr. Isabella Anderson-Wagner PhD

"*A game is being played at the extremity of this infinite distance where heads or tails will turn up. What will you wager? According to reason, you can do neither the one thing nor the other; according to reason, you can defend neither of the propositions. Do not, then, reprove for error those who have made a choice; for you know nothing about it.*"

Blaise Pascal

UNIVERSITY OF VIENNA

Abstract

Faculty of Psychology

Department of Cognition, Emotion, and Methods of Psychology

Doctor rer. nat.

Theory and Application of Bayesian Inference in Social Neuroscience

by Lukas Leopold LENGERSDORFF

Recent years have seen a renaissance of Bayesian thinking in psychology and neuroscience. In this cumulative dissertation, I explore how the use of Bayesian inference can deepen our understanding of human social behavior and its neural bases. The thesis is structured in two parts, reflecting the dual focus of my research.

In the first part, I investigate how Bayesian methods can enhance inference in standard experimental designs. In an fMRI study on the effects of violent video games on empathy, I demonstrate how Bayesian analyses allow statements about the presence and absence of experimental effects. In a theoretical paper, I then critique common misunderstandings about statistical power, and discuss how Bayesian and classical frequentist perspectives differ in what they allow us to infer from data.

In the second part, I focus on the use of Bayesian cognitive modelling, specifically reinforcement learning models, in social neuroscience. I first present an experimental fMRI study investigating the neurocomputational basis of prosocial learning and decision-making. Based on insights from this empirical work, I then discuss common pitfalls and best practices for the application of reinforcement learning models. Finally, I develop an efficient approach to the Bayesian testing of constrained hypotheses, and demonstrate how it can be used to test scientific expectations about human reinforcement learning.

Together, these five contributions illustrate how Bayesian modelling can serve as a pragmatic tool for data analysis, and how Bayesian thinking can broaden our understanding of the scientific process itself. My aim is to bridge empirical neuroscience and statistical theory, and to illustrate the potential of a diverse methodological toolkit.

Acknowledgements

Writing this thesis was quite the journey, and I am grateful to, and grateful for, all the people who helped me along the way.

First and foremost I would like to thank my supervisor, Claus Lamm. You believed in me when I started this PhD, and you continued to believe in me even when things got rough. This thesis would not have been possible without your wisdom, your generosity, and your patience. *Danke!*

Moreover, I am very grateful to my co-supervisors, Isabella Wagner and Predrag Petrovic, who supported me immensely, especially during the beginning of my PhD, and helped me find my footing in the world of neuroscience.

I would also like to say a very special *Dankjewel* to my supervisor during my research stay in the Netherlands, Eric-Jan Wagenmakers. EJ, you made me feel at home in Amsterdam, both as a mathematical psychologist and as a human. I will always cherish the occasions we just sat in your office and brainstormed whatever statistical idea came to our minds.

Of course, *the real thesis are the friends we made along the way.* I would never have been able to do this without the support and friendship of my PhD buddies, Magdalena Boch and Helena Hartmann. Leni, at the time I am writing this, you are on (tenure) track to become a PI and professor, and I cannot think of anybody who deserves this more than you. Thank you for all the times you let me verbally ventilate my frustrations in your office, and helped me through the roughest patches of this PhD. The same goes for you, Heli: Thank you for all the support and kindness you showed me (and for bringing a bit of German order into my chaotic Austrian day).

I am very thankful to have done my PhD in the SCAN Unit, a lab filled with great minds and great hearts. Thank you for giving me a place to grow as a scientist. Also, I am very sorry for all the lab meetings I hijacked by obsessing over some minor statistical issue!

Finally, I'd like to use this space to send lots of *Bussis* and a paper-mediated hug to my little family, Veronika and Pauline. Veronika, thank you for supporting me through all of this - even when supporting me meant telling hyper-caffeinated 5am me to shut the laptop and finally go to bed. Pauline, the majority of my ideas for Chapter 6 of this thesis came to me while I carried you across the Zaanse Schans and the beach at Castricum aan Zee. *Inspired the main intellectual contribution through nap walks* is not yet a point in the CRediT Contributor Role Taxonomy, but it definitely should be! Also, I am sorry for all the afternoons I spent writing this thesis instead of baking sand cake with you. Love, Papa!

Contents

Für Pauline

Chapter 1

General Introduction

It is fair to say that Bayesian inference is "in". Especially in psychology and neuroscience, Bayesian methods are widely regarded as a promising alternative to classical frequentist approaches such as p-values and significance tests (e.g., Benjamin et al., 2018; Coventry & Bartlett, 2024; Dienes, 2016; Dienes & Mclatchie, 2018; Keysers et al., 2020; McElreath, 2018; Wagenmakers et al., 2011, 2018). In the wake of the replication crisis (Bakker et al., 2012; Open Science Collaboration, 2015), Bayesian statistics have even been proposed as a means to conduct more replicable, robust, and reliable research — in short, to make science "better" (e.g. Benjamin et al. 2018; Romero and Sprenger 2021; Wagenmakers et al. 2011, see Colling and Szűcs 2021 for a critical review). This development has been particularly prominent in social neuroscience, where researchers now have access to numerous tools for fitting Bayesian models to social behavior (e.g. Ahn et al., 2017; Daunizeau et al., 2014; C. D. Mathys et al., 2014).

However, using Bayesian inference also comes with many challenges, most of which stem from its relatively recent introduction into researchers' toolkits. Whereas there is broad consensus on the "correct" ways to use frequentist inference, the adequate application of Bayesian inference in social neuroscience is much less codified. Should competing hypotheses be tested with Bayes factors (Kass & Raftery, 1995; Keysers et al., 2020) as the Bayesian "equivalent" of p-values, or should inference rely on directly interpreting the posterior distribution of parameters (e.g., Gelman et al., 2013, 2020; Kruschke, 2011)? Should default priors (Berger & Mortera, 1999; Rouder et al., 2012; Wetzels et al., 2012) be used, or should priors incorporate researchers' domain knowledge and expectations (Stefan et al., 2022)? And in what way can Bayesian inference be integrated into a field still dominated by frequentist statistics?

In this thesis, I explore how Bayesian methods can deepen our understanding of human social behavior and its neural bases. My research focus is twofold: First, I examine how Bayesian inference can enhance and complement the conclusions drawn from standard experimental designs. Second, I demonstrate how Bayesian cognitive modeling can be used to investigate the computational mechanisms underlying human social behavior, with a particular focus on reinforcement learning in the social domain.

In the following, I briefly introduce the logic behind Bayesian inference and modeling, and provide an overview of the current state of Bayesian methods in social neuroscience. I then outline the five papers that constitute the two parts of this thesis. Finally, I offer some personal context on the development of this dissertation project.

1.1 The Principles of Bayesian Inference

1.1.1 Bayes' Theorem

At the heart of Bayesian inference lies the equation famously known as Bayes' theorem (Bayes, 1763):

$$p(A|B) = \frac{p(B|A)}{p(B)} \cdot p(A), \tag{1.1}$$

where A and B are *events* that are assumed to be random, $p(A)$ is the probability of A without additional consideration of B, and $p(A|B)$ is the *conditional* probability of A given that B has already occured. Equation 1.1 is a standard result of probability theory relating the conditional probabilities of random events to their marginal probabilities, used widely across all disciplines of statistics, and by frequentists and Bayesians alike[1].

What is now widely known as *Bayesian inference*, however, is the application of Bayes' theorem to *update scientific beliefs in hypotheses after observing data*. For example, assume that we observe data related to some research hypothesis $\mathcal{H}$. Then using Bayes' theorem in the spirit of Bayesian inference gives us

$$p(\mathcal{H}|\text{data}) = \frac{p(\text{data}|\mathcal{H})}{p(\text{data})} \cdot p(\mathcal{H}), \tag{1.2}$$

where the terms are defined as follows:

- $p(\mathcal{H})$ is the *prior probability* of hypothesis $\mathcal{H}$, representing our initial belief in the hypothesis,
- $p(\mathcal{H}|\text{data})$ is the *posterior probability* of $\mathcal{H}$, representing our belief in the hypothesis **after** we have observed the data,
- $p(\text{data}|\mathcal{H})$ is the probability of observing such data under the assumption that the hypothesis holds, also known as the *marginal likelihood*, and
- $p(\text{data})$ is the total probability of observing such data, regardless of whether the hypothesis holds or not.

In words, equation 1.2 says that our belief in a hypothesis, which is represented by its probability, increases if the observed data were more likely to be observed if the hypothesis were true than if the hypothesis were false.

In practical applications, the data will almost always be assumed to be generated by a parametric *statistical model* specified in accordance with the research hypothesis $\mathcal{H}$. Thus, the data are assumed to follow the probability distribution

$$p(\text{data} \mid \theta, \mathcal{H}),$$

where θ is a *parameter* further governing the model's behavior. Importantly, different hypotheses may be represented by different statistical models with possibly different sets of parameters. For example, one hypothesis $\mathcal{H}_0$ may propose that measurements from two groups follow a normal distribution with the same mean and variance, another hypothesis $\mathcal{H}_1$ may propose that the two groups have different means, and

[1] A popular "non-Bayesian" example, which many readers will know from their introductory Statistics course, is the use of Bayes' theorem to calculate the probability that a patient has some disease, given that a test for that disease gave a positive result.

another hypothesis $\mathcal{H}_2$ may even assume that the measurements follow a distribution other than the normal distribution.

In contrast to frequentist inference, Bayesian statistical models further necessitate the formulation of *prior distributions* over the models' parameters,

$$\theta \sim \pi(\theta \mid \mathcal{H}).$$

The prior π may be understood to encode initial knowledge or beliefs about plausible values of the parameter. Whereas in equation 1.2 we applied Bayes' theorem on the level of models/hypotheses, we now apply it on the level of parameters within a model to derive the *posterior distribution* of θ:

$$\pi(\theta|\text{data}, \mathcal{H}) = \frac{p(\text{data}|\theta, \mathcal{H})}{p(\text{data}|\mathcal{H})} \cdot p(\theta|\mathcal{H}). \tag{1.3}$$

Here, the term $p(\text{data}|\mathcal{H})$ is the marginal likelihood of the data under the hypothesis, which also appeared in the hypothesis-level version of Bayes' theorem (equation 1.2). It is given by the integral

$$p(\text{data}|\mathcal{H}) = \int_\Theta p(\text{data}|\theta, \mathcal{H})\ \pi(\theta|\mathcal{H})\ d\theta, \tag{1.4}$$

and represents the probability of observing such data under the hypothesized model *without* reference to a specific value of θ. As we will see in section 1.1.3, the marginal likelihood is a central concept in one approach towards Bayesian inference.

1.1.2 Fitting Models to Data

In theoretical terms, fitting a Bayesian model to data means obtaining the posterior distribution (equation 1.3) by conditioning the prior on the observed data. In practice, however, this is a non-trivial task, as the integral that defines the marginal likelihood (equation 1.4) can only be solved analytically in a small fraction of cases (Chib & Jeliazkov, 2001; Wasserman, 2000).

The modern solution is not to compute the posterior distribution directly, but to approximate it by drawing samples. This has become feasible thanks to the development of Markov Chain Monte Carlo (MCMC) algorithms (Robert et al., 1999), which allow sampling from a distribution when its probability density function is known up to a constant. In the case of the posterior, this unknown constant is the marginal likelihood.

1.1.3 Modes of Bayesian Inference

Once a model has been fitted to data, there are different ways to draw inferences from it, depending on the analytical goals and philosophical perspective. In practice, Bayesian inference in psychology and neuroscience often follows one of two major pathways: one centered on the use of Bayes factors, and another that avoids them in favor of estimation and prediction.

Using Bayes Factors

Consider the case that we have two different hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, corresponding to two different models. After observing data, we may ask two questions:

1. Under which of these two models were the data more likely?
2. How much do the data change the posterior probability of one hypothesis relative to the other?

Both of these questions are answered by the Bayes factor (BF; Kass & Raftery, 1995), which is defined as the ratio of the marginal likelihoods of the models:

$$BF_{1,0} = \frac{p(\text{data}|H_1)}{p(\text{data}|H_0)}. \tag{1.5}$$

It is immediately clear how the BF answers question 1: If it is larger than 1, then the data were more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$; if it is smaller than 1, the data were more likely under $\mathcal{H}_0$. But the BF also tells us how much the data change the odds of one hypothesis holding true relative to the other (question 2). To see this, we simply form the ratio of the posterior probabilities (as defined in equation 1.2) of the two hypotheses:

$$\begin{aligned} \frac{p(\mathcal{H}_1|\text{data})}{p(\mathcal{H}_0|\text{data})} &= \frac{\frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data})} \cdot p(\mathcal{H}_1)}{\frac{p(\text{data}|\mathcal{H}_0)}{p(\text{data})} \cdot p(\mathcal{H}_0)} \\ &= \underbrace{\frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data}|\mathcal{H}_0)}}_{BF} \cdot \frac{p\left(\mathcal{H}_1\right)}{p\left(\mathcal{H}_0\right)}. \end{aligned} \tag{1.6}$$

Thus, the BF may be understood as an indicator of the evidence for one hypothesis over the other given by the data, independent of the priors we assign to these hypotheses.

The BF has become a highly popular metric, probably due to a combination of several factors. For one, its interpretation is highly intuitive, and its use has been suggested by several influential methodologists (e.g. Dienes, 2016; Keysers et al., 2020; Wagenmakers et al., 2018). Moreover, it may be understood as a Bayesian equivalent to the p-value, as it allows to make inferential decisions based on a single number. This is particularly accomodating to researchers used to the conventions of frequentist inference, such as binary decisions between "significant" and "not significant" results. Accordingly, even though the BF is a continuous measure of evidence, it has become increasingly common to use conventional thresholds marking BFs as "anecdotal", "substantial", or "strong" evidence for hypotheses (e.g. Kass & Raftery, 1995; Keysers et al., 2020). Additionally, perhaps one of the most popular uses of the BF is to obtain evidence for the *absence* of an experimental effect - something that is not easily achieved within the frequentist framework (e.g. Dienes, 2014; Dienes et al., 2018; Keysers et al., 2020). Lastly, its implementation in popular and easily accessible software packages such as JASP (JASP Team, 2025) make the BF a low-threshold inferential tool for researchers without extensive training in Bayesian methodology.

Using the BF also comes with challenges and limitations, though. One challenge is computational - while MCMC sampling foregoes the need to calculate the marginal likelihood to arrive at posterior samples, it is exactly this quantity that is needed to calculate the BF. In most cases, the involved integral (equation 1.4) cannot be solved analytically, and must therefore be estimated by numerical means. Several algorithms

have been developed to approximate the marginal likelihood efficiently (e.g. Gronau et al., 2017), but a precise estimate often still comes with high computational costs.

Perhaps the most important downside of the BF is that it tends to be highly sensitive to the choice of prior distributions assigned to model parameters (Hoijtink, 2022; Kruschke, 2011). This is unproblematic whenever researchers are able to formulate informative priors that reflect existing knowledge and expectations about the parameters (e.g. Vanpaemel, 2010). However, when researchers have not enough information to specify informative priors, sensitivity of test results to arbitrary choices may be seen as undesirable. Solutions to this issue are the use of default priors, which have been designed to lead to sensible BFs in the majority of testing cases (e.g. Berger & Mortera, 1999; Rouder et al., 2012; Wetzels et al., 2012), as well as sensitivity analyses that gauge how much the choice of different priors influence the results (Liu & Aitkin, 2008).

Avoiding Bayes Factors: Estimation and Prediction

Not all Bayesian inference is based on Bayes factors. Indeed, several influential methodologists discourage the use of the BF (e.g. Gelman & Rubin, 1995; Gelman et al., 2013) and recommend alternative approaches to inference. In general, these approaches put less emphasis on hypothesis testing, and more emphasis on directly examining the posterior distribution of parameters of interest, as well as assessing the accuracy of the model when predicting new data.

The posterior distribution provides complete information about the most likely values of a parameter given the observed data, and can therefore be used for estimation. Point estimates of the parameter can be obtained by calculating summary statistics of the posterior distribution, such as the posterior mean, median, or mode. Uncertainty regarding the parameter can be quantified with *credible intervals*, which contain the true parameter with a certain, prespecified probability (e.g., 95%). Credible intervals may also serve as the basis for testing a null hypothesis against an alternative, for example through a Region-of-Practical-Equivalence (ROPE; Kruschke, 2011) analysis. Here, the researcher specifies an interval of parameter values that are "practically equivalent" to the value specified by the null hypothesis, and assesses whether or not the credible interval lies outside or inside of this interval.

In practice, however, researchers do not only consider a single model, but rather need to select the most sensible model out of a larger pool of candidates. In approaches that avoid the use of BFs for model comparison, this selection is typically performed by assessing the predictive accuracy of models when applied to new data. Popular methods to achieve this are cross-validation, as well as the use of information criteria (see Vehtari & Ojanen, 2012, for an overview). Additionally, researchers may perform posterior predictive checks (PPC; Gelman et al., 2013; Lynch & Western, 2004) to assess the absolute fit of the model to the observed data. Here, the fitted model is used to simulate new data, to assess whether the model can generate key features of the actually observed data.

In general, these approaches provide a promising alternative for researchers who are skeptical of Bayes factors. However, they are arguably less accessible to those without extensive experience in Bayesian inference. Easy-to-use interfaces such as JASP are lacking, and users are often required to work with probabilistic programming languages like Stan (Carpenter et al., 2017). Moreover, influential proponents of this perspective have emphasized that the goal of Bayesian inference should not

be hypothesis testing at all, but rather the iterative development and refinement of predictive models (Gelman & Rubin, 1995; Gelman et al., 2020). While this is a defensible position from a theoretical point of view, it can be at odds with the expectations in fields such as psychology and neuroscience, where researchers are typically required to provide clear, decisive results to support their claims.

1.2 Bayesian Methods in Social Neuroscience

Social neuroscience is a field in which enormous amounts of data are fitted to highly complex models in order to explain phenomena characterized by profound uncertainty — namely, social behavior. It is therefore perhaps not surprising that Bayesian inference has gained considerable traction among researchers of the field, both as an analytical tool and as a framework for reasoning about the brain itself.

Within contemporary social neuroscience, Bayesian inference plays three distinct roles. First, it is used to enhance and complement conclusions drawn from classical frequentist statistics. Second, it enables the formulation and estimation of complex computational models of cognition that cannot easily be accommodated by traditional statistical methods. Third, it serves as a theoretical framework for modeling cognition itself. In such accounts, the brain is assumed to engage in a form of belief updating akin to Bayesian inference, in order to learn from and adapt to its environment.

1.2.1 Bayesian Inference as a Complement to Frequentist Inference

In social neuroscience, as in many other fields, the increasing availability of Bayesian inference due to computational advances has led to its re-evaluation as an alternative to, and possible replacement of, frequentist significance testing. Numerous publications outline the problems associated with the misinterpretation and misuse of the p-value, and offer the adoption of Bayesian inference as a solution (e.g. Coventry & Bartlett, 2024; Dienes & Mclatchie, 2018; Keysers et al., 2020). Often, these accounts suggest that Bayesian inference allows statements about what researchers are often "actually" interested in: the probability of the tested hypothesis being true, given the data (whereas the often misunderstood p-value only gives the probability of the observed data, and more extreme data, under the assumption that the null hypothesis is true).

Perhaps the most popular feature of Bayesian inference is its ability to provide evidence for both the alternative and the null hypothesis, that is, for the presence *and* absence of experimental effects (Dienes, 2014; Dienes et al., 2018; Keysers et al., 2020). This is in contrast to the p-value which, in itself, cannot distinguish between insignificant results that are obtained because the tested effect does not exist, and insignificant results that are obtained because the conducted test was not sensitive enough (but for frequentist alternatives to assert the null hypothesis, see e.g. Lakens et al., 2018).

1.2.2 Bayesian Estimation of Cognitive Models

Classical statistical tools, such as t-tests, ANOVAs, or correlations, allow social neuroscientists to assess how often certain social behaviors occur, and how their intensity varies across situations or populations. However, these methods provide little insight into the underlying cognitive processes that give rise to such behaviors. As a result, researchers have increasingly turned to mathematical *cognitive models* that

aim to approximate the mental mechanisms underpinning social phenomena. Such models explain individual behavior as the result of a dynamic interaction between environmental stimuli and internal cognitive variables, such as values, beliefs, or expectations, that are not directly observable but must be inferred from behavior. These variables are represented numerically, and the structure of the cognitive process is captured by parameterized mathematical formulas (Forstmann & Wagenmakers, 2015; Lewandowsky & Farrell, 2010).

Bayesian modeling has become an invaluable tool in the formulation, estimation, and comparison of cognitive models. While such models can in principle be estimated using frequentist methods (e.g., through maximum likelihood estimation), doing so is often computationally challenging and inflexible, particularly for models with latent variables, complex hierarchical structures, and non-linear relationships between parameters, experimental variables, and observed behavior. In contrast, Bayesian modeling allows the straightforward estimation of nearly any cognitive model, as long as it can be expressed in a probabilistic programming language such as STAN (Carpenter et al., 2017), Turing (Ge et al., 2018), Edward2 (Tran et al., 2018), or NumPyro (Phan et al., 2019).

It should be noted, however, that researchers employing Bayesian cognitive modeling rarely adopt the full Bayesian philosophy of inference. In practice, it is common for model estimation and selection to rely on Bayesian methods, while subsequent statistical analyses revert to traditional frequentist approaches — for instance, by testing correlations between estimated parameters and participant traits using p-values, or analyzing fMRI data via conventional significance-based whole-brain analyses (see e.g., Crawley et al., 2020; Kutlikova et al., 2023; Mikus et al., 2022, 2023; Su et al., 2025; Zhang & Gläscher, 2019). This concurrent use of Bayesian and frequentist techniques reflects a pragmatic stance toward inference, likely driven by both publication norms and the limited availability or accessibility of Bayesian alternatives for common neuroscientific procedures (e.g., mass-univariate fMRI analysis).

In social neuroscience, perhaps the most widely used cognitive model is the reinforcement learning (RL) model, which allows researchers to probe into the mechanisms underlying learning and decision-making in the social domain (Lockwood & Klein-Flügge, 2020). RL modeling has been used to examine learning 'for' others (learning in order to make decisions that are beneficial to others; e.g. Lockwood et al., 2016, 2019), learning 'from' others (learning through observing behaviors of other individuals; e.g. Behrens et al., 2008; Burke et al., 2010; Hill et al., 2017; Lindström, Bellander, et al., 2019; Lindström, Golkar, et al., 2019; Suzuki et al., 2012; Zhang & Gläscher, 2019), and learning 'about' others (acquiring information about the behaviors and beliefs of others Hampton et al., 2008; Lockwood et al., 2018; Will et al., 2017; Yoon et al., 2018; Zhu et al., 2012).[2]. Many studies employing RL modeling investigate differences in learning and decision making between social and non-social situations, which can be reflected by differences in RL parameters. Particularly valuable for neuroscientific research, RL modeling also allows the extraction of trial-wise indicators of the underlying cognitive processes, such as subjective values and prediction errors. These can then be used as covariates in fMRI analyses, to identify common and distinct neural bases for social and non-social learning.

[2]This taxonomy of studies into learning 'for', 'from' and 'about' others was developed by my colleague Lei Zhang for the co-authored tutorial paper which is printed as Chapter 4 of this thesis

1.2.3 Bayesian Models of Human Cognition

Given the effectiveness of Bayesian inference as an analytical tool, it is not surprising that neuroscientists have also adopted it as a theoretical framework for how the brain itself makes sense of the world (Bottemanne, 2024). In these "meta-Bayesian models"[3] of cognition, the brain is assumed to integrate sensory input (the "data") with prior beliefs and knowledge (the "prior") to form an updated internal representation of the world (the "posterior"). This view aligns with both scientific findings and everyday observations suggesting that perception is rarely purely objective, but instead shaped by expectations and prior experience (De Lange et al., 2018). One of the most influential of these theories is *Active inference* (see Moutoussis et al., 2014; Pezzulo et al., 2024, for review). Here, it is assumed that humans (as well as other sentient agents) constantly try to minimize discrepancies between their environment and their internal representation of the same environment. They may do so either by updating their internal representation when it does not align with the environment (i.e., perception and learning), or by changing their environment to be more similar to their representation (i.e., action).

When applied to social behavior, Bayesian models of cognition often take on additional layers of complexity. In such accounts, an agent's internal model includes representations of others' mental states (their beliefs, goals, and intentions) who in turn may be modeling the agent's own mental states, and so on. For example, in cooperative interactions, agents may simultaneously track a partner's current intentions, the reliability of their advice, and the volatility of their behavior over time. Hierarchical Bayesian models can formalize these interacting layers of uncertainty and belief updating (e.g. C. D. Mathys et al., 2014). Complex, multilayered models such as these exemplify the value of Bayesian thinking for social neuroscience, both as tool and theory.

1.3 Chapter Outline

With this cumulative thesis, I wish to illustrate the wide applicability of Bayesian inference in neuroscientific research, and to support researchers who wish to make it part of their own methodological toolkit. The papers contained in this thesis are organized into two parts, representing the dual focus of my empirical and theoretical research.

In **Part I**, I focus on the application of Bayesian methods to enhance and complement the conclusions drawn from standard experimental designs. Chapter 2 presents an fMRI study on the effects of violent video games on empathy and emotional reactivity. Given the inconclusive literature on this controversial topic, this study provided the ideal opportunity to illustrate how Bayesian inference may provide evidence for both the presence and the absence of experimental effects. In Chapter 3, I discuss common misconceptions about statistical power, and provide guidelines towards a principled assessment of results from low-powered tests. Within my thesis, this paper serves as

[3] Admittedly, terminology can become confusing in this area of neuroscience. The terms "Bayesian model" and "Bayesian inference" are sometimes used to describe analytical tools employed by the *researcher*, and other times to describe cognitive strategies attributed to the *subject* or their brain. Bayesian models *as tools* may be used to address research questions entirely unrelated to Bayesian theories of cognition, and hypotheses derived from such theories can be tested using non-Bayesian methods. Typically, though, researchers who view Bayesian inference as a helpful framework for understanding the brain are also inclined to use it in their analyses.

an example on how Bayesian thinking can deepen our understanding of the scientific process itself.

Part II explores the use of Bayesian cognitive modeling in social neuroscience, with a particular focus on reinforcement learning models. Chapter 4 takes the form of a tutorial paper on the principled application of reinforcement learning models, and discusses common pitfalls in the interpretation of modeling results. In Chapter 5, I present an empirical study on the neurocomputational bases of prosocial learning and decision making. Here, I used Bayesian hierarchical reinforcement learning models to gauge how humans learn to protect others from pain. Finally, in Chapter 6, I develop Sampling in Constrained Space (SICS), a method for the efficient testing of equality and inequality constraints in all kinds of statistical models. Moreover, I demonstrate its efficiency and wide generalizability by testing whether or not human reinforcement learning is characterized by decreasing learning rates.

1.4 Development of this Thesis

It is fair to say that this thesis grew non-linearly. During my PhD, I applied Bayesian modeling on a wide variety of different research questions, both empirical and theoretical. Although I hope that the final product stands as a self-contained demonstration of the utility of Bayesian thinking in social neuroscience, I deem it helpful to shortly explain the history of the papers collected here.

I originally set out to investigate the effects of violent video games on empathy and emotional reactivity, as reported in Chapter 2. The design of that study necessitated two experimental sessions and, due to the small sizes of reported violent video game effects, a sample that was much bigger than typical for fMRI studies. Given this large investment, my supervisor Claus Lamm and I found it sensible to add a prosocial learning task to the experimental paradigm. This way, we could use the data from the first experimental session to also investigate the neurocomputational underpinnings of prosocial learning and decision making (Chapter 5).

It was during the analyses for this study that I first used Bayesian inference and cognitive modeling, in particular Reinforcement Learning (RL) modeling. However, I soon encountered several challenges. For one, specifying, estimating, and comparing different RL models was a very complex endeavour, and I found relatively little guidance on this topic in the published literature. Moreover, I identified what I thought to be common misunderstandings regarding the interpretation of modeling results (in particular, the interpretation of high learning rates as indicating "good" or "fast" learning). In discussions with my supervisor and my colleagues Lei Zhang and Nace Mikus, we soon realized the need for a standalone publication that would describe common misunderstandings and pitfalls in RL modeling, as well as suggest best practices. We subsequently wrote and published the tutorial paper that is now Chapter 4 of this thesis.

Due to the large sample size and complex experimental design, collecting the data for the violent video game study took over two years. During that time (and beyond), I extensively studied Bayesian modeling, as well as topics of statistical inference and Philosophy of Science. Moreover, I "followed my passion" and formally enrolled in the Statistics programme of the University of Vienna, from which I graduated while working on this thesis. One question that particularly intrigued me was in what ways Bayesian statistics can enhance inference in research fields that predominantly use

"classical" frequentist statistics, and in what way the concurrent use of both statistical approaches can be justified. In line with this, I started questioning the common review practice of critiquing study results based on the power of the conducted tests. I found it very difficult to justify this practice, as it entailed an in my view inconsistent judgement of frequentist analyses on the basis of Bayesian ideas. I discussed these ideas with my supervisor, and we decided to elaborate on this issue in a theoretical paper (Chapter 3).

Driven by my interest in Bayesian inference and cognitive modeling, I then completed a one-year research stay at the Bayesian Methodology group lead by Eric-Jan Wagenmakers at the University of Amsterdam, the Netherlands. There, I started developing a novel way to estimate and compare constrained statistical models, which may also be used to test scientific hypotheses about human reinforcement learning. Prof. Wagenmakers connected me with his colleague, Maarten Marsman, with whom I subsequently wrote Chapter 6 of this thesis.

However, as my interest expanded from social neuroscience to statistical methodology, I realized that my main contribution to the field did not lie in the investigation of the effects of violent video games on social behavior, but rather in the theoretical development and empirical application of Bayesian methods to neuroscientific research questions. After consulting with my supervisor, I therefore changed the original topic of my thesis to the one presented here.

Part I

Enhancing Standard Inference

Chapter 2

Neuroimaging and behavioral evidence that violent video games exert no negative effect on human empathy for pain and emotional reactivity to violence

Abstract

Influential accounts claim that violent video games (VVG) decrease players' emotional empathy by desensitizing them to both virtual and real-life violence. However, scientific evidence for this claim is inconclusive and controversially debated. To assess the causal effect of VVGs on the behavioral and neural correlates of empathy and emotional reactivity to violence, we conducted a prospective experimental study using functional magnetic resonance imaging (fMRI). We recruited eighty-nine male participants without prior VVG experience. Over the course of two weeks, participants played either a highly violent video game, or a non-violent version of the same game. Before and after this period, participants completed an fMRI experiment with paradigms measuring their empathy for pain and emotional reactivity to violent images. Applying a Bayesian analysis approach throughout enabled us to find substantial evidence for the absence of an effect of VVGs on the behavioral and neural correlates of empathy. Moreover, participants in the VVG group were not desensitized to images of real-world violence. These results imply that short and controlled exposure to VVGs does not numb empathy nor the responses to real-world violence. We discuss the implications of our findings regarding the potential and limitations of experimental research on the causal effects of VVGs. While VVGs might not have a discernible effect on the investigated subpopulation within our carefully controlled experimental setting, our results cannot preclude that effects could be found in settings with higher ecological validity, in vulnerable subpopulations, or after more extensive VVG play.

2.1 Introduction

Video games have evolved into one of the most popular forms of entertainment. In Europe, 25% of the population report playing video games weekly, and especially young

This chapter is published as **Lengersdorff, L. L.**, Wagner, I. C., Mittmann, G., Sastre-Yagüe, D., Lüttig, A., Olsson, A., Petrovic, P., & Lamm, C. (2023). Neuroimaging and behavioral evidence that violent video games exert no negative effect on human empathy for pain and emotional reactivity to violence. *eLife*, *12*, e84951. https://doi.org/10.7554/eLife.84951

adults spend much time in these "virtual worlds" (IPSOS MediaCT, 2012). Many popular games contain high levels of violent imagery, with the killing or hurting of other characters being deeply engrained in the gameplay (Gentile et al., 2004; Krantz et al., 2017). Many recent studies have investigated whether such violent video games (VVGs) have adverse effects on real world social behavior and empathy (Anderson et al., 2010). According to the influential General Aggression Model (Bushman & Anderson, 2002), VVGs should decrease the players' empathy for the pain of others by desensitizing them to both virtual and real violence. Such desensitizing effects should in turn be reflected by decreased activity in brain areas underpinning empathy, such as the anterior insula (AI) and the anterior midcingulate cortex (aMCC) (Lamm et al., 2011, 2019). However, the evidence for this prediction is mixed. While some studies found that playing VVGs leads to emotional desensitization on the behavioral and neural level (Arriaga et al., 2011; Bartholow et al., 2006; Carnagey et al., 2007; Engelhardt et al., 2011; Staude-Müller et al., 2008), other studies failed to reveal such effects (Gao et al., 2017; Kühn et al., 2019; Szycik, Mohammadi, Hake, et al., 2017; Szycik, Mohammadi, Münte, & te Wildt, 2017). Conflicting results are also found on the level of systematic reviews (de Vrieze, 2018; Mathur & VanderWeele, 2019). Several meta-analyses suggest that VVGs exert small, yet consistent adverse effects on aggression and empathy (Anderson et al., 2010; Calvert et al., 2017; Greitemeyer & Mügge, 2014; Mathur & VanderWeele, 2019; Prescott et al., 2018). Other researchers contest these results, claiming that results are a product of selective reporting and biased analyses (Ferguson & Kilburn, 2010; Hilgard, Engelhardt, & Rouder, 2017).

A key question is whether VVGs are causally responsible for low empathy, or whether less empathic individuals are more likely to play VVGs (Bushman & Anderson, 2015; Ferguson & Kilburn, 2010). Many studies have been quasi-experimental in nature, comparing the empathic responses of participants who habitually play VVGs with those of participants without VVG experience (Bartholow et al., 2005, 2006; Gentile et al., 2016; Krahe et al., 2011). Such designs provide limited information on the direction of the causal link between VVGs and decreased empathy. The existing experimental studies have nearly always used VVGs as an experimental manipulation shortly before measuring the outcomes of interest (Arriaga et al., 2011; Bushman & Anderson, 2009; Carnagey et al., 2007; Engelhardt et al., 2011; Guo et al., 2013; Staude-Müller et al., 2008). While these studies consistently report evidence for a desensitizing effect of violent games, they cannot disentangle the immediate effects of VVG play from those that have a persistent, long-term impact on individuals. Immediate VVG effects may encompass a wide range of processes, such as priming (Bushman, 1998), as well as stress-like responses such as increases in active fear and aggressive behaviors (Fanselow, 1994; Mobbs et al., 2007, 2009) that include generally increased sympathetic activity, release of stress hormones, heightened activation of involved brain structures, and cognitive-affective responses (e.g., deep reflection on the seen content, and changes in emotions and mood). Such responses can persist on a time-scale of minutes to hours after aversive events such as VVG exposure, and have been shown to negatively affect social behavior (Nitschke et al., 2022). It is important to distinguish these immediate effects from longer-term adaptations that occur over days or weeks, such as habituation or memory consolidation processes. The General Aggression Model predicts that the repeated exposure to violence in the positive emotional context of videogames leads to the gradual extinction of aversive reactions, resulting in the long-term desensitization of players to real-world violence (Bushman & Anderson, 2009).

It is therefore essential to conduct experimental studies that can disentangle the

long- and short-term effects of VVGs in participants without prior VVG experience. One first such study was conducted by Kühn et al. (2019) who found no significant effects of VVGs on empathy and its neural correlates. While this study was an important starting point, four important design features limited its conclusions. First, the researchers used very dissimilar games in the experimental group versus the control group, restricting the comparability of the two conditions. Second, while the participants of the experimental group were asked to play the violent game Grand Theft Auto V (Rockstar Studios) for thirty minutes per day over two months, the authors did not control the degree to which participants actually played the game. Third, the authors did not control that participants actually committed violent acts within the game, as the game offers a large amount of gameplay without violent content. Fourth, the absence of significant results was interpreted as evidence for the absence of VVG effects. However, the authors did not report the results of equivalence tests (Lakens et al., 2018) or Bayesian hypothesis tests (Keysers et al., 2020)that would support such claims conclusively (Hilgard, Engelhardt, Bartholow, & Rouder, 2017). In view of the many conflicting results reported by experimental research and even meta-analyses (de Vrieze, 2018; Mathur & VanderWeele, 2019), clearly differentiating between "absence of evidence" and "evidence of absence" is particularly important.

To test possible causal effects of VVGs on empathy and its neural correlates, we conducted an experimental prospective study, which addressed each of these limitations. Eighty-nine male participants with little to no prior VVG experience repeatedly played a modified version of Grand Theft Auto V over the course of two weeks. Participants in the experimental group played a highly violent version of the game and were tasked to kill as many other characters as possible. Participants of the control group played a version of the same game from which all violent content was removed, and were asked to perform a non-violent task (taking photographs of other characters). Before and after this gaming period, participants completed an fMRI session during which we measured the behavioral and neural correlates of empathy for pain and emotional reactivity to violent images (see Figure 2.1 and Methods: experimental fMRI sessions for details). We used Bayesian hypothesis tests to assess whether there were negative effects of VVGs on participants' empathic behavior and neural responses. Hypothesis tests were performed by means of the Bayes Factor (BF; Kass & Raftery, 1995). We followed the convention to report a BF > 3 as evidence for the alternative hypothesis, a BF < 1/3 as evidence for the null hypothesis, and a BF in the interval [1/3, 3] as inconclusive evidence for either hypothesis (Kass & Raftery, 1995; Keysers et al., 2020). We would like to emphasize, though, that the Bayes factor provides an easily interpretable continuous quantification of the evidence for and against hypotheses, and that a strict categorization of BFs into evidence for and against hypotheses is not necessary. Our aim was to provide conclusive evidence on the question whether VVGs can desensitize humans to the plight of others or not, within our carefully balanced experimental model.

2.2 Results

2.2.1 Behavioral Data

Descriptive statistics of gaming behavior

Forty-Five participants took part as part of the experimental group, and 44 participants as part of the control group. On average, participants of the experimental group killed 2844.7 characters (SD = 993.9, Median = 2820, Minimum = 441, Maximum =

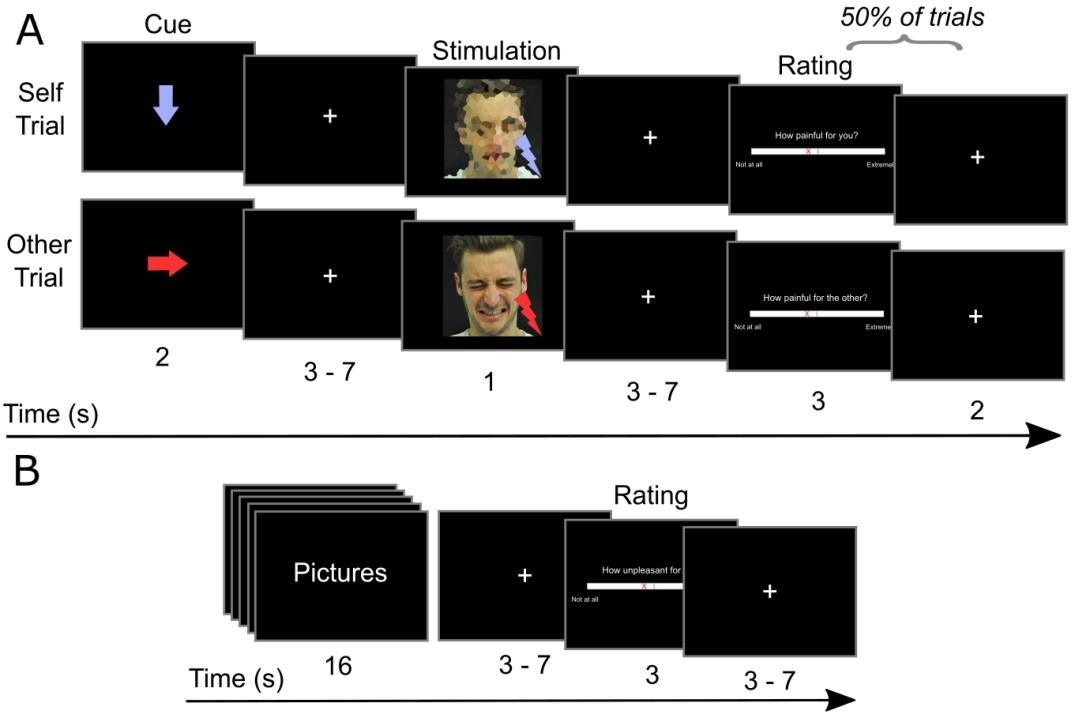

FIGURE 2.1: Schematic depiction of the experimental tasks. **A)** Empathy-for-pain task. In trials of the Self condition, participants passively received electrical stimuli. In the Other condition, participants experienced how another person (a confederate) received electrical stimuli. The stimuli were either painful or not painful. In the cue phase, an arrow indicated the recipient (downwards: Self; right: Other) and the intensity (blue: not painful; red: painful) of the next stimulus. In the stimulation phase, the stimulus was delivered. After half of the trials, participants were asked to rate the last stimulus. The confederate depicted has given informed consent that his photograph can be published. **B)** Emotional reactivity task. Participants were presented pictures with different content (violent or neutral) and different context (real or game context). After observing a block of pictures, participants rated their current unpleasantness on a visual analoge scale from 0 to 100.

6815). Participants of the control group took an average of 3055.3 pictures of other characters (SD = 1307.5, Median = 3026, Minimum = 441, Maximum = 6815). Thus, as was the aim of our experimental design, each participant of the experimental group was exposed to a substantial number of violent acts in the video game.

Empathy for pain

To test our central hypothesis, we investigated if participants who played the violent video game showed decreased empathy for pain on the behavioral level. We analyzed the ratings obtained during the empathy-for-pain task with a hierarchical Bayesian censored regression model. We modelled fixed effects for the experimental factors *Group* (non-violent vs. violent gaming, coded as -1 and 1), *Time* (pre vs. post gaming sessions, coded as -1 and 1), and *Intensity* (non-painful vs. painful stimulation of the confederate, coded as -1 and 1), as well as all interactions between these factors. See Methods, Data analysis for more details.

The posterior means of fixed effect parameters are listed in Table 2.1 for both painfulness and unpleasantness ratings. As a manipulation check, we first tested whether painful stimuli led to increased painfulness and unpleasantness ratings, compared to non-painful stimuli. For both kinds of ratings, this test revealed very strong evidence (BF > 100) for an effect of Intensity, indicating that our paradigm was able to induce empathic responses in participants (see Figure 2.2, A and B). The posterior mean of the regression parameter β of the factor Intensity was 27.86 for Painfulness ratings, and 17.48 for Unpleasantness ratings. Given our used factor coding, this means that the average difference in ratings between painful and non-painful stimuli was 2*27.86 = 55.72 points of the 100-point VAS for Painfulness ratings, and 2*17.48 = 55.72 points for Unpleasantness ratings.

We found evidence for the absence of a VVG effect on the painfulness ratings. Comparing a model where the fixed effect of *Group*Time*Intensity* could be negative to a model where the effect was set to zero resulted in a BF of 0.324. This means that the observed ratings were about 3.1 times more likely under the null hypothesis of no VVG effect than under the alternative hypothesis. When estimated without restrictions, the posterior mean of β for the interaction *Group*Session*Intensity* was -0.78. Given our factor codings, this means that the quantity $[\text{rating}_{\text{Pain}} - \text{rating}_{\text{No Pain}}]_{\text{Session 2}} - [\text{rating}_{\text{Pain}} - \text{rating}_{\text{No Pain}}]_{\text{Session 1}}$ (thus, the baseline corrected empathic response) was on average 1.56 points smaller in the experimental group than in the control group, on the 100-point VAS scale. However, note that the Bayesian hypothesis test suggests that a model with this interaction restricted to zero provides a better explanation of the data.

For the unpleasantness ratings, evidence for absence of a VVG effect was substantial. With a BF of 0.130, the observed data were about 7.7 times more likely under the null hypothesis of no VVG effect than under the alternative hypothesis. The posterior mean of β for the interaction *Group*Session*Intensity* was -0.45. Given our factor codings, this means that the quantity $[\text{rating}_{\text{Pain}} - \text{rating}_{\text{No Pain}}]_{\text{Session 2}} - [\text{rating}_{\text{Pain}} - \text{rating}_{\text{No Pain}}]_{\text{Session 1}}$ was on average 0.9 points smaller in the experimental group than in the control group. However, note again that the Bayesian hypothesis test suggests that a model without this interaction provides a better explanation of the data.

In summary, the behavioral data suggest that violent video game play as implemented in this study has no effect on either type of empathy rating.

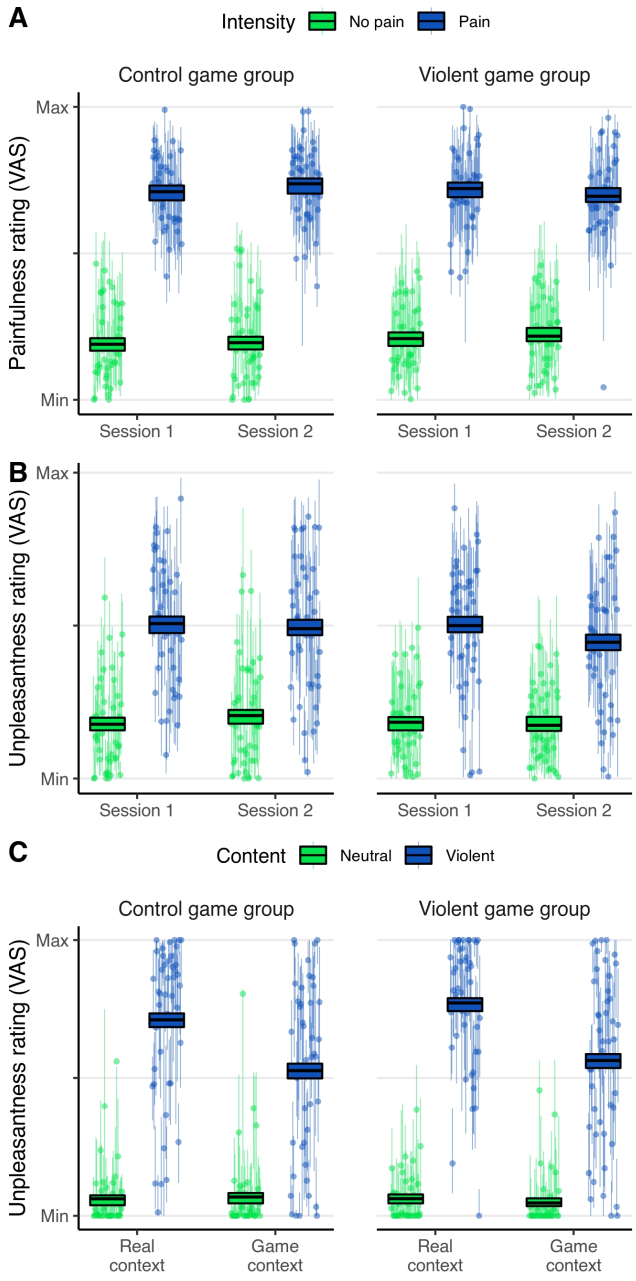

FIGURE 2.2: Behavioral results. Depicted are participants' ratings during the empathy-for-pain task (A and B) and the emotional reactivity task (C). Ratings were given on a visual analog scale (range: 0-100). **A) Empathy for pain, Painfulness rating.** Question text: "How painful for the other?". **B) Empathy for pain, Unpleasantness rating.** Question text: "How unpleasant for yourself?". Note that an apparent trend towards a three-way interaction Group*Session*Intensity is not supported by the respective Bayesian hypothesis test (BF = 0.130, Table 1). **C) Emotional reactivity, unpleasantness rating.** Question text: "How unpleasant?". Boxes: the middle line marks the group mean of participant ratings in the respective condition; the box represents the 95% credible interval of the posterior predictive distribution of mean ratings. Dots depict the individual mean ratings of participants, lines depict the 95% credible interval of the posterior predictive distribution of mean ratings of single participants.

TABLE 2.1: Posterior parameter means of models for ratings in the empathy-for-pain task. Dependent variable: empathy ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Session: first session = -1, second session = 1; Intensity: non-painful = -1, painful = 1. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

Fixed effect	β	**95% Credible Interval**		**Bayes Factor**
Painfulness ratings				
Group	0.66	-1.14	2.46	0.127
Session	0.42	-0.31	1.16	0.072
Intensity	27.86	25.32	30.29	>100
Group*Session	-0.48	-1.24	0.25	0.102
Group*Intensity	-1.12	-3.48	1.32	0.207
Session*Intensity	-0.26	-1.38	0.83	0.069
Group*Session*Intensity	-0.78	-1.87	0.32	0.324†
Unpleasantness ratings				
Group	-1.13	-4.83	2.56	0.251
Session	-0.63	-1.94	0.71	0.141
Intensity	17.48	14.93	20.11	>100
Group*Session	-0.93	-2.33	0.40	0.254
Group*Intensity	-0.95	-3.63	1.68	0.197
Session*Intensity	-1.20	-2.12	-0.30	0.996
Group*Session*Intensity	-0.45	-1.36	0.49	0.130†

Emotional reactivity

Next, we investigated whether playing the violent video game desensitized participants towards depictions of violence. We again used a hierarchical Bayesian censored regression model, and included fixed effects for the experimental factors *Group* (non-violent vs. violent gaming, coded as -1 and 1), *Content* (neutral vs. violent, coded as -1 and 1), and *Context* (real vs. game, coded as -1 and 1).

The posterior means of fixed effect parameters of this model are listed in Table 2.2. As a manipulation check, we first tested whether participants experienced more unpleasantness in the emotional reactivity task while observing violent pictures compared to neutral pictures. We found very strong evidence (BF > 100) for this hypothesis, indicating that our paradigm was successful in inducing unpleasantness by violent imagery. The posterior mean of the regression parameter β of the factor Content was 37.08. This means that the average difference in ratings between violent and neutral stimuli was 74.16 points of the 100-point VAS. The unpleasantness ratings are depicted in Figure 2.2 C.

Further, we found substantial evidence for the absence of a desensitizing VVG effect. Comparing a model where the fixed effect of *Group*Content* could be negative to a model where the effect was set to zero resulted in a BF of 0.151. Thus, participants of the violent game group did not show a decreased emotional response towards depictions of real and game violence. Moreover, testing the fixed effect of *Group*Content*Context* resulted in a BF of 0.094, indicating that there was also no

desensitizing effect that was specific to depictions of game violence. When estimated without restrictions, the regression parameters associated with both interactions were positive, β = 2.28 for Group*Content, and β = 0.33 for *Group*Content*Context*. This means that, ostensibly, participants in the experimental group had a very weak tendency to rate violent images as more unpleasant than participants in the control group, contrary to expectations. However, note again that the Bayesian hypothesis test suggests that a model without these interactions provides a better explanation of the data. In summary, the behavioral data suggest that playing the violent video game did not emotionally desensitize participants towards violent images.

TABLE 2.2: Posterior parameter estimates of models for ratings in the emotional reactivity task. Dependent variable: unpleasantness ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Content: neutral = -1, violent = 1; Context: real = -1, game = 1. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

Fixed effect	β	**95% Credible Interval**		**Bayes Factor**
Group	1.26	-2.78	5.44	0.349
Content	37.08	32.98	41.48	>100
Context	-7.24	-9.15	-5.39	>100
Group*Content	2.28	-1.92	6.52	0.151†
Group*Context	-1.23	-3.01	0.47	0.306
Content*Context	-5.36	-7.39	-3.34	>100
Group*Content*Context	0.33	-1.58	2.10	0.094†

2.2.2 fMRI Data

Empathy for pain

We next analyzed the fMRI data collected during the empathy-for-pain task. To define our ROIs, we first performed whole-brain general linear model (GLM) analysis of the data of the first fMRI session. Our contrast of interest [*Other Pain - Other No Pain*] compared brain activity when the confederate experienced painful stimulation to activity when the confederate experienced only non-painful stimulation (see Methods, Data analysis for details). This revealed significant clusters in our a priori defined brain areas of interest, aMCC and bilateral AI, as well as in other areas, including the left supramarginal gyrus and the right angular gyrus (see Figure 2.3A, and Appendix 2.A for detailed results).

Subsequently, we performed Bayesian linear mixed effects analyses on the data extracted from the ROIs (aMCC, left AI, right AI). See Methods, Data analysis for details. We compared models where the fixed effect of *Group*Time*Intensity* could be negative to a model where the effect was set to zero. For responses in the *Cue* phase (where participants were informed whether the other person would receive a painful or a non-painful stimulus), we obtained the following BFs: $BF_{aMCC} = 0.402$; $BF_{\text{left AI}} = 0.547$; $BF_{\text{right AI}} = 0.190$. For responses in the *Stimulation* phase (where participants observed the other person receiving the stimulus), we obtained the following BFs:

$BF_{aMCC} = 0.176; BF_{left\ AI} = 0.143; BF_{right\ AI} = 0.434$. See Appendix 2.B, Table 2.B.1 for posterior distributions and Bayes factors of all model parameters.

In summary, we found weak to moderate evidence for the absence of an effect of playing the violent video game on participants' brain activity while they observed another person in pain.

Emotional reactivity

Our next analysis concerned the fMRI data coming from the emotional reactivity task. To define our ROIs, we computed the contrast [*Violent – Neutral*], comparing brain activity during observation of violent images to brain activity during observation of images with neutral content (see Methods, Data analysis for details). This revealed significant clusters in one of our a priori areas of interest, the bilateral amygdala, as well as several other regions, such as the bilateral fusiform gyrus and the bilateral precentral gyrus (see Figure 2.3B, and Appendix 2.A for detailed results). However, we found no significant clusters in the other brain regions of interest, the aMCC or the bilateral AI. Therefore, we restricted our subsequent ROI analysis to the amygdala.

We performed Bayesian linear mixed effects analyses on the data extracted from the amygdala. Methods, Data analysis for details. First, we compared a model where the fixed effect of *Group*Content* could be negative to a model where the effect was set to zero. This resulted in a BF of 0.324 for the left amygdala, and a BF of 0.338 for the right amygdala, indicating absence of an effect in both ROIs. Next, we tested the fixed effect of *Group*Content*Context*. With a BF of 0.205 for the left amygdala, and 0.163 for the right amygdala, this analysis also indicated the absence of an effect. See Appendix 2.B, Table 2.B.2 for posterior distributions and Bayes factors of all model parameters.

In summary, the data suggest that playing the violent video game did not lead to a dampened brain response to images of violence in neither real nor gaming contexts.

2.2.3 Post-hoc Analyses

Sample comparability

We constrained our sample to young adult (18 – 35 years) males who had minimal prior exposure to VVGs in general, and who had not played the game used in the study before. However, given the great popularity of VVGs among young adult males, it is also possible that this constrained our sample to a subpopulation that is less susceptible to desensitization effects to begin with. Therefore, we tested whether the subpopulation from which we drew our sample exhibited higher levels of trait empathy than the general population. To achieve this, we compared the trait empathy levels of our sample, as measured by the Questionnaire for Cognitive and Affective Empathy (QCAE; Reniers et al., 2011), to those of a control sample of 18-35 year old males who were not preselected for minimal VVG use. See Methods: Data analysis: Post-hoc analyses for more details.

The results are depicted in Table 2.3. For all subdimensions, Bayesian t-tests provided moderate to substantial evidence for the hypothesis that there is no difference between the two groups ($BF < 1/3$). Thus, our exploratory analysis suggests that our inclusion criterion of minimal VVG exposure did not result in a preselection of individuals with extraordinarily high levels of empathy.

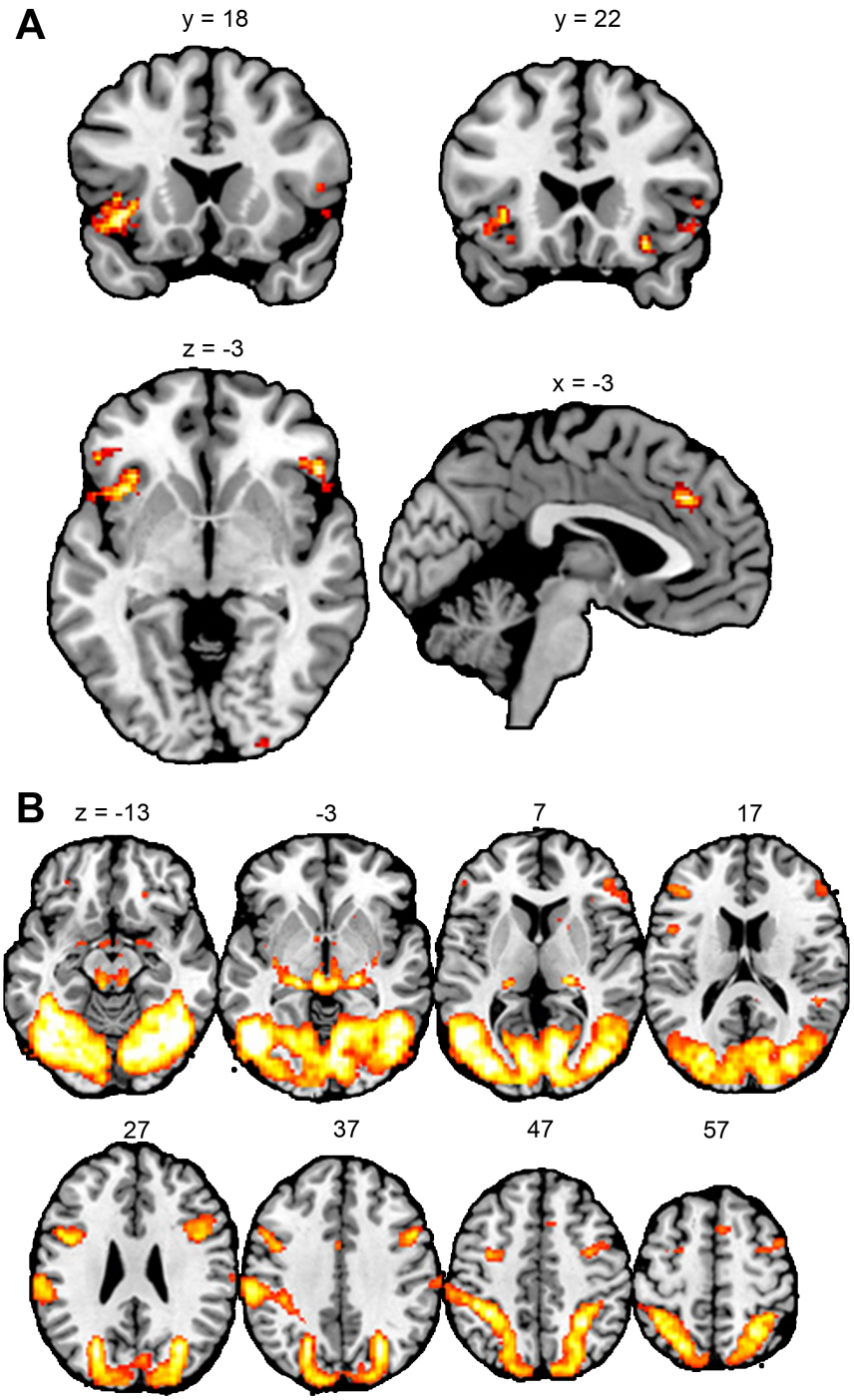

FIGURE 2.3: Results of the whole-brain analyses for region-of-interest definition. **A) Empathy-for-pain task.** Clusters represent areas where brain activity was increased when the confederate received a painful electrical stimulus, compared to a non-painful stimulus. **B) Emotional reactivity task.** Clusters represent areas where brain activity was increased during the observation of violent images, compared to neutral images. All results $p < 0.05$ FWE-corrected. This figure was made with the software MRICron (https://www.nitrc.org/projects/mricron).

TABLE 2.3: Comparison of trait empathy levels between experimental group and control group. Bayes Factors were derived from comparing a model where the mean difference VVG group – Control group was positive to a model where it was restricted to zero (one-sided Bayesian t-test).

	VVG group (N = 83)		Control group (N = 132)			
QCAE Subdimension	**Mean**	**SD**	**Mean**	**SD**	**t**	**Bayes Factor**
Perspective Taking	1.93	0.43	2.01	0.51	-1.189	0.074
Online Simulation	1.96	0.41	1.92	0.47	0.588	0.259
Emotional Contagion	1.58	0.48	1.63	0.55	-0.659	0.098
Peripheral Responsivity	1.58	0.58	1.58	0.62	0.021	0.155
Proximal Responsivity	1.67	0.58	1.79	0.53	-1.538	0.063

Test-retest reliabilities

In this study, we measured a number of behavioral and neural correlates in two experimental sessions – once before the exposure to the violent video game or the control game, once after. Thus, the test-retest reliability (i.e., the correlation between the two measurements of a variable) is of interest, as this informs us about the relative stability of our outcome variables of interest. This also affects the statistical power of our performed tests (see next section). For analysis details, see Methods: Data analysis: Post-hoc analyses. We found that the test-retest reliability of our behavioral measures of empathy (i.e., participants' ratings) was high to very high (Painfulness ratings: ρ = 0.768, 95% credible interval = [0.613, 0.879]; Unpleasantness ratings: ρ = 0.905, 95% credible interval = [0.813, 0.967]). However, we observed very low test-retest reliability for our neural measurements of empathy (AMCC signal: ρ = -0.013, 95% credible interval = [-0.420, 0.402]; left AI signal, ρ = -0.001, 95% credible interval = [-0.423, 0.414]; right AI signal, ρ = 0.027, 95% credible interval = [-0.377, 0.416]).

Bayesian design analyses

We based our sample size on the results of a power analysis designed for the frequentist inference framework (see Methods: Power analysis). However, as we ultimately based our inference on Bayes factor tests, the theoretical long-term behavior of these tests, given our sample size and expected effect size, is of interest. This also informs us about the effect sizes that could realistically have been detected using our sample size. Therefore, we conducted a post-hoc Bayes Factor design analysis by means of a Monte Carlo simulation experiment (Schönbrodt & Wagenmakers, 2018). See Methods: Data analysis: Post-hoc analyses for analysis details. It is of particular importance to note that the diagnosticity of hypothesis tests involving repeated measurements does also depend on the correlation between the repeated measures, i.e. the test-retest reliability.

The results are presented in Table 2.4. In summary, the simulation experiment suggested that our behavioral analyses, for which test-retest reliabilities were high, were well enough powered to differentiate between the absence and presence of a medium-to-small effect of d = 0.3. Note that this effect size is smaller than the lower bound of effect size estimates reported in the meta-analysis of Anderson et al. (2010), which was d = 0.345. For smaller effects, such as d = 0.2, the a priori power of our behavioral

TABLE 2.4: Results of the Bayes factor design analysis. Depicted are the estimated probabilities of inferential decisions for each dependent variable and assumed true effect size d. Inc.: Inconclusive evidence, no decision. H0: evidence for the null hypothesis. H1: evidence for the alternative hypothesis. ρ = correlation between repeated measurements, i.e. test-retest reliability. The estimated probabilities of correct decisions (evidence for H0 when d = 0.0, evidence for H1 when d > 0.0) are marked in **bold**.

	Empathy									**Emotional Reactivity**		
	Painfulness ratings (ρ = 0.75)			Unpleasantness ratings (ρ = 0.90)			Neural response (ρ = 0)			Unpleasantness ratings (only second session)		
Effect size	**Inc.**	**H0**	**H1**	**Inc.**	**H0**	**H1**	**Inc.**	**H0**	**H1**	**Inc.**	**H0**	**H1**
d = 0.0	0.29	**0.69**	0.02	0.29	**0.69**	0.02	0.30	**0.69**	0.02	0.30	**0.68**	0.02
d = 0.2	0.58	0.20	**0.23**	0.44	0.05	**0.50**	0.49	0.43	**0.08**	0.55	0.33	**0.13**
d = 0.3	0.48	0.06	**0.46**	0.14	0.00	**0.85**	0.56	0.30	**0.14**	0.57	0.18	**0.25**
d = 0.4	0.28	0.02	**0.71**	0.02	0.00	**0.98**	0.57	0.21	**0.22**	0.50	0.08	**0.42**

analyses was not optimal, as it would have been likely that we would have obtained an inconclusive result ($1/3 < BF < 3$) even in the presence of a true effect of that size. However, given that we obtained evidence for the null hypothesis ($BF < 1/3$) in all relevant Bayes Factor tests on our behavioral data, our results speak strongly against the presence of such an effect.

Regarding our neural analyses, given the low correlation between repeated measurements (i.e. test-retest reliability), the Bayesian power of our fMRI analyses should be regarded as low. Taken alone, we would not consider them convincing evidence against the presence of a VVG effect. However, together with our behavioral results, they suggest that VVG effects, if they exist, can be expected to be very small.

Cross-task correlations

Given that we measured empathy for pain and emotional reactivity in the same subjects, our data also allowed us to investigate the relationships between these two phenomena. For this, we calculated the correlations between the behavioral and neural measurements of our outcome variables. The results are presented Table 2.5. We can observe that for our behavioral measures, cross-task correlations were substantial (r = .227 - .280, with all credible intervals not covering zero). However, we could observe no substantial cross-task correlations for our neural measures, or across neural and behavioral indicators.

TABLE 2.5: Cross-task correlations. Above diagonal: posterior means of correlations. Below diagonal: 95% credible intervals of correlations. Unpl. = Unpleasantness. Emo.Reac. = Emotional Reactivity. AMCC = anterior midcingulate cortex. AI = anterior insula. Amy = Amygdala. l. = left. r. = right.

		Emp: Pain	**Emp: Unpl.**	**ER: Unpl.**	**Emp: AMCC**	**Emp: lAI**	**Emp: rAI**	**ER: lAmy**	**ER: rAmy**
	Empathy: Pain		.630	.280	.015	.043	.133	-.037	-.037
Behavior	**Empathy: Unpl.**	(.544, .708)		.227	.039	.084	.211	.058	.096
	Emo. Reac.: Unpl.	(.191, .366)	(.130, .323)		-.010	.028	-.028	.060	.043
	Empathy: AMCC	(-.178, .215)	(-.170, .245)	(-.212, .190)		.055	.069	.040	.050
	Empathy: l. AI	(-.151, .235)	(-.118, .276)	(-.167, .214)	(-.201, .285)		.104	.019	.021
Neural	**Empathy: r. AI**	(-.056, .310)	(-.009, .393)	(-.205, .153)	(-.156, .289)	(-.146, .323)		.080	.085
	Emo. Reac.: l. Amy	(-.178, .096)	(-.079, .195)	(-.080, .202)	(-.164, .241)	(-.172, .213)	(-.124, .266)		.507
	Emo. Reac.: r. Amy	(-.173, .098)	(-.042, .230)	(-.095, .180)	(-.150, .251)	(-.171, .207)	(-.117, .271)	(.370, .628)	

2.3 Discussion

Influential theories of media violence predict that the repeated playing of VVGs results in decreased empathy for pain due to a desensitization to real world violence (Anderson et al., 2010; Bushman & Anderson, 2002). Here, we report evidence against this hypothesis in relation to our specific setting. We found that participants who repeatedly played a highly violent game for seven hours over the course of two weeks did not show decreased empathy for another person's pain or decreased responses to violent imagery.

Our findings contrast with several earlier studies that found a negative relationship between playing VVGs and empathic responses to violence. Importantly, the majority of these studies were quasi-experimental in nature, and therefore provide only limited evidence for a putative causal effect of violent gaming (Bartholow et al., 2005, 2006; Gentile et al., 2016; Krahe et al., 2011). Moreover, the few experimental studies that exist implemented designs investigating short-term carry over effects, as they had exposed participants to virtual violence rather immediately before measurements of their outcome variables of interest (Arriaga et al., 2011; Bushman & Anderson, 2009; Carnagey et al., 2007; Engelhardt et al., 2011; Guo et al., 2013; Staude-Müller et al., 2008). Together with the study of Kühn et al. (2019), our study is one of the first to investigate persistent effects of VVGs in participants without prior experience with them, enabling a clear assessment of the causality of VVG effects. Importantly, our study was designed to address several limitations of the study of Kühn et al.: We strictly controlled the amount of virtual violence actually experienced by participants, and used a non-violent version of the same game in the control condition; moreover, we applied a Bayesian analytical approach, which, together with our comparatively large sample size, enabled us from the outset to distinguish "absence of evidence" from "evidence of absence" of VVG effects. This approach yields consistent evidence from both behavioral and neural data that violent video games, to the extent and characteristics played in our interventional design, are not causally responsible for a persistent lack of empathy or emotional desensitization to violence.

Despite the aforementioned strengths of our study, we also need to address several limitations. Our experimental design ensured that participants of the experimental group were exposed to a substantial amount of violent gameplay during gaming sessions (each participant „killed“ an average of 2000 other characters in a graphically violent way). However, the overall exposure to virtual violence was still very low when compared to the amount that is possible in the everyday life of typical VVG players. During our experiment, participants played for 7 hours over the course of two weeks. However, habitual gamers can play an average of 16 hours in the same time frame (Clement, 2021; Statista Research Department, 2022). Our results cannot preclude that longer and more intense exposure to VVGs could have negative causal effects on empathy. In particular, adolescents and children as well as persons with specific neuropsychiatric traits might be especially susceptible to long-term changes due to increased brain plasticity. However, empirically testing higher levels of violence with the same degree of control as realized in our study would reach the limits of practical feasibility. We thus believe that our results provide an important perspective on the size of VVG effects that could realistically be expected in experimental research.

To increase experimental control, we restricted our sample to young adult males who had minimal prior exposure to violent video games. It is possible that, due

to this strict preselection criterion, our sample was drawn from a subpopulation that is particularly resistant to desensitization. An exploratory analysis provided strong evidence that our selection criterion did not result in particularly high levels of trait empathy in our sample, though. However, we cannot preclude that our sample was particularly resistant to VVG effects due to other, untested characteristics. Further research is needed to assess if our results generalize to samples with other characteristics that may be more representative for the general population.

To maximize the amount of violence that participants would be exposed to (and commit) in the game, we restricted the game's objective to killing other characters, and incentivized this behavior with monetary rewards. This might have reduced the ecological validity of our operationalization of gaming, and it is possible that bigger effects could be seen when violent gameplay is more internally motivated, i.e. individuals who want to play the game may be differently affected than those that have merely accepted to be part of an experiment. Still, our results provide valid evidence that the mere exposure to virtual violence for seven hours over two weeks is not sufficient to decrease empathy. It should be noted that there are few studies that connect laboratory-based experimental investigations of empathy and emotional reactivity to real-world behavior and its measures. There are indications, however, that neuroscientifc empathy measures similar to the ones used here predict individual social behavior (e.g., donation, helping or care-based behavior; Ashar et al., 2017; Hartmann et al., 2022; Hein et al., 2010; Tomova et al., 2017), and that they are also validated by their predictivity of mental or preclinical disorders characterized by deficits in empathy (Bird et al., 2010; Lamm et al., 2016). That said, it is obvious that future research is needed that bridges and integrates laboratory and field-based measures and approaches, in order to inform us how changes (or their absence) in neural responses induced by VVG play are connected to real-life social emotions and behaviors (see Stijovic et al., 2023, for a recent example illustrating, in the domain of social isolation research, how a combined lab- and field based study can be directly informed by prior laboratory-based neuroscience findings).

Our study was designed to reliably detect an effect size of d = 0.3, an effect even smaller than the lower estimate for VVG effects on empathy reported in Anderson et al. (2010). Our results provide substantial evidence that effects of this magnitude are not present in settings similar to our experimental design. These arguments notwithstanding, it needs to be noted that future studies with higher power may detect still smaller effects. Considering the high prevalence of VVG, even such small effects could be of high societal relevance (Funder & Ozer, 2019). For now, based on the current design and data, we can conclude that experimental long-term VVG effects on empathy are unlikely to be as large as previously reported.

It may be argued that the empathy for pain paradigm and the associated behavioral and neural responses are so robust and resistant to changes by external factors that this may explain the lack of evidence for the effects of VVG play. This argument however would contradict a wealth of findings illustrating malleability of empathic responses using this and related designs, including with placebo analgesia (Rütgen, Seidel, Riečanský, & Lamm, 2015; Rütgen, Seidel, Silani, et al., 2015; Rütgen et al., 2021), an intervention that usually shows low to moderate effect sizes as well (see e.g. Hein & Singer, 2008; Jauniaux et al., 2019; Lamm et al., 2019, for review).

Lastly, and somewhat surprisingly, we found that the test-retest reliability of our neural covariates of empathy for pain were close to zero for all investigated ROIs. Knowing that an individual's neural empathic response (BOLD activity for seeing

somebody else in pain vs. in no pain) was above or below average in the first session provides little to no information about their relative response in the second session. To the best of our knowledge, our study was the first one to present the empathy for pain paradigm to the same sample of participants after a longer time frame. Thus, this surprising result provides valuable information on the limitations of this task respectively the neural measurements acquired in it, and certainly demands further research to investigate the factors influencing fMRI reliability (see also Elliott et al., 2020; Kragel et al., 2021). We would like to emphasize, though, that a high test-retest reliability is not a precondition for the valid testing of group-level effects. For a group-level effect to be testable, it is only necessary that the mean of the dependent variable is consistently affected by the independent variable. It is not necessary that participants who show an above average level in the DV in one session also show an above average level in the second session, and vice versa. Otherwise, there would also be no point in independent-sample designs. Indeed, it has recently been discussed that highly robust cognitive tasks are bound to exhibit low test-retest reliability, as robust tasks are often characterized by low inter-individual variation, and thus leave only little variance that can be explained by participant traits (Hedge et al., 2018). However, it must also be noted that low reliability does lead to lower power of repeated measures designs. As discussed above, the low reliability of the measured neural responses has resulted in suboptimal power of our tests on fMRI data.

In summary, our findings stand in contrast to claims that posit the playing of violent games as an essential factor for explaining decreases in empathy. If this is shown to generalize to when people play more often and over longer periods, the desensitization to violence described in prior reports using quasi-experimental designs might have been caused by third and pre-existing factors, such as education, socio-economic status, or mental health issues (DeCamp & Ferguson, 2016; Lemmens & Bushman, 2006; Shao & Wang, 2019; Tortolero et al., 2014). Together with similar findings (Kühn et al., 2019), our results point out the limits to which VVGs can be held responsible for lacks of empathy, at least in highly controlled experimental settings that last for the two weeks of play implemented here. This is not to say, though, that there is no point in further investigating the complex relationships between violent media use and adverse social behavior. We propose that the design and analysis approach of the present study could act a reference of how future studies should be conducted, in order to increase the stringency and robustness of research in this domain. Together with our findings, such studies will aid in resolving the scientific controversy regarding the negative effects of violent video games (de Vrieze, 2018; Mathur & VanderWeele, 2019), and contribute to a deeper understanding of the interplay between violent media and emotion.

2.4 Methods

2.4.1 Power analysis

We planned to collect data from 90 participants. We derived this sample size from a power analysis based on VVG effect sizes reported in the meta-analysis of Anderson et al. (2010). The authors estimated the size of the negative VVG effect on empathy/desensitization to be r = 0.194, 95% CI = [0.170, 0.217], which corresponds to Cohen's d = 0.396, 95% CI = [0.345, 0.445], representing a small-to-medium effect. We chose d = 0.300 as the minimum effect size for which we wanted to achieve a power of 0.80, to ensure that we would have enough power even if the reported effect size

was overestimated. Note that thus, the effect size we used was even smaller than the lower bound reported in Anderson et al. (2010). We performed the power analysis using the software Gpower 3.1.9.2. (Faul et al., 2007), calculating the required sample size to achieve a power of 0.8 for the interaction in a 2-by-2 within-between design ANOVA, assuming a medium correlation of 0.5 between repeated measures, and using the conventional alpha error level of 0.05. This resulted in a required sample size of 90. Using such a sample size, the achieved power for the effect size reported in (Anderson et al., 2010), as well as its lower and upper bound, was as follows: for d = 0.345, achieved power = 0.901; for d = 0.396, achieved power = 0.960; for d = 0.445, achieved power = 0.986.

Please note that while this power analysis was based on a frequentist analysis framework, we are reporting Bayesian analyses here. However, we considered this power analysis to be a sensible benchmark for the sample size needed to answer our research questions. See Results: Post-hoc analyses: Bayesian design analysis for a Bayesian design analysis that provides more information on the size of effects that could be detected with our sample size using Bayesian analyses.

2.4.2 Participants

In total, 97 participants completed the first experimental session. Of these, eight participants dropped out of the study (six before the first video game sessions; two after, of which one was from the experimental group and one from the control group). We thus acquired complete datasets from 89 participants.

To control for previous VVG exposure, we only included individuals that had not played VVGs at least twelve months before testing, and had not played the video game Grand Theft Auto V before. We did this to avoid a possible ceiling effect: participants who had already played these games before might already have been desensitized too much for our experimental VVG exposure to show any effect, therefore reducing sensitivity. We tested only male participants, as more males than females play violent video games regularly (Gentile et al., 2004; Krahé & Möller, 2004; Padilla-Walker et al., 2010). Moreover, males have been shown to be more easily influenced by violent media (Bartholow & Anderson, 2002; Bettencourt & Kernahan, 1997). To further increase homogeneity of the sample, we restricted the age range of possible participants to 18-35 years. Additional inclusion criteria were no history of neurological or psychatric disorders or drug abuse, and standard inclusion criteria for MRI measurements. Participants were recruited through online advertisements and received a financial compensation of €145 for participating in all experimental sessions. A performance-linked bonus of up to €35 acted as an additional incentive during the game sessions. The study was approved by the ethics committee of the Medical University of Vienna (decision number 1258/2017). The confederate depicted in Figure 2.1 A has given informed consent that his photograph may be used for this publication.

2.4.3 Overall study design

Participants were randomly assigned to the violent game group or the control game group. Participants first completed a pretest fMRI session, during which they performed an experimental task designed to measure empathy for pain. Then, over the course of two weeks, participants of the violent game group repeatedly played a violent video game, while the control game group played a non-violent version of the

same game. Subsequently, both groups completed the posttest fMRI session. Here, participants performed the empathy-for-pain paradigm again, and also completed a task designed to measure emotional reactivity to violent pictures.

2.4.4 Experimental fMRI sessions

Confederate

To facilitate empathic responses during the experimental tasks, participants completed the experimental session together with a male confederate. The confederate acted as if he were a second participant of the experiment. This deception was maintained until the end of the last experimental session, at which point participants were debriefed.

Pain calibration

The empathy-for-pain paradigm included the administration of painful but tolerable stimuli. The physical pain was induced via a well-established procedure (e.g., Rütgen, Seidel, Silani, et al., 2015). Electrical stimuli were produced by a Digitimer DS5 stimulator (Digitimer Ltd, Clinical & Biomedical Research Instruments, United Kingdom) and delivered by electrodes placed on the dorsum of the left hand. Subjective pain thresholds were determined using a standardized calibration procedure. The participant received short (500 ms) stimuli of increasing intensity and was asked to rate pain intensity on a numeric scale (0 = "not perceptible"; 1 = "perceptible, but not painful", 3 = "a little painful", 5 = "moderately painful", 7 = "very painful", 9 = "extremely painful, highest tolerable pain"). The average intensities of stimuli rated as 1 and 7 were then chosen as the intensities of the non-painful and painful stimulation conditions during the empathy-for-pain task.

Empathy-for-pain paradigm

We used a well-established paradigm to measure participants' empathic responses (Hartmann et al., 2021; Rütgen, Seidel, Riečanský, & Lamm, 2015; Rütgen, Seidel, Silani, et al., 2015; Singer et al., 2004). Participants either received electric stimuli themselves (*Self* condition), or saw images of the confederate indicating that he was currently receiving electric stimulation (*Other* condition). The stimuli were either painful (*Pain* condition) or perceptible but not painful (*No Pain* condition). The timeline of the task is illustrated in Figure 2.1A. At the start of each trial, a downwards or rightwards arrow (presented for 2 s) indicated whether the next stimulus would be delivered to the participant or the confederate, respectively (*Cue* phase). Red and blue arrows indicated painful and non-painful stimulation, respectively. After a jittered interval [3-7 s], the stimulus was delivered (*Stimulation* phase). In the Self condition, the participant received the electrical stimulus (0.5 s), and saw a pixelated photograph (1 s). In the Other condition, the participant saw a photograph of the confederate with a neutral or painful facial expression. After half of the trials, participants rated the last stimulus on a 100-step visual analog scale (VAS). In the Self condition, participants rated how painful the last stimulus was for themselves. In the Other condition, participants rated how painful the stimulus was for the confederate (other-oriented painfulness rating), and how unpleasant it was for themselves to observe the confederate receiving the stimulus (self-oriented unpleasantness rating). In total, there were 64 trials, with 16 trials per condition (Self Pain, Self No Pain, Other Pain, Other No Pain). Conditions were presented in a pseudorandomized order.

The task was presented using COGENT (http://www.vislab.ucl.ac.uk/cogent.php), implemented in MATLAB 2017b (The MathWorks Inc., Natick, MA, USA). The total task duration was approx. 20 min.

Emotional reactivity paradigm

To investigate emotional reactivity to violent images, we used an affective picture paradigm (Olofsson et al., 2008; Petrovic et al., 2005). Participants were shown pictures of either neutral or violent content (factor *Content*). Additionally, the pictures depicted either real scenes, or scenes taken from the video game participants played during the gaming sessions (factor *Context*). Real pictures were taken from the International Affective Pictures System (IAPS; Lang et al., 2005). Game pictures were matched to IAPS pictures in terms of content, valence and arousal (see Appendix 2.C).

The sequence of events of the task is illustrated in Figure2.1B. Each block consisted of five pictures of the same condition, presented for 3 seconds each, and with a short interval of 0.2 seconds between pictures. After a jittered interval [3-7 s] participants rated how unpleasant they felt on a 100-step VAS. In total, participants saw 16 blocks of pictures, with 4 blocks per condition (Neutral Real, Neutral Game, Violent Real, Violent Game). The task was presented using COGENT, and total task duration was approx. 5 min. To avoid that participants formed expectations about the purpose of the study early on, participants completed this task only in the second fMRI session.

MRI data acquisition

MRI data were acquired with a 3T Siemens Skyra MRI system (Siemens Medical, Erlangen, Germany) and a 32-channel head coil. Blood oxygen level dependent (BOLD) functional imaging was performed using a multiband-accelerated echoplanar imaging sequence with the following parameters: Echo time (TE): 34 ms; Repetition time (TR): 1200ms; flip angle: 66°; interleaved ascending acquisition; 52 axial slices coplanar to the connecting line between anterior and posterior commissure; multiband acceleration factor 4, resulting in 13 excitations per TR; field-of-view: 192 ×192 × 124.8 mm, matrix size: 96 × 96, voxel size: 2×2×2 mm, interslice gap 0.4 mm. Structural images were acquired using a magnetization-prepared rapid gradient-echo sequence with the following parameters: TE = 2.43 ms; TR = 2300 ms; 208 sagittal slices; field-of-view: 256×256×166 mm; voxel size: 0.8×0.8×0.8 mm. To correct functional images for inhomogeneities of the magnetic field, field map images were acquired using a double echo gradient echo sequence with the following parameters: TE1/TE2: 4.92/7.38 ms; TR = 400 ms; flip angle: 60°; 36 axial slices with the same orientation as the functional images; field-of-view: 220×220×138 mm; matrix size: 128×128×36; voxel size: 1.72×1.72×3.85 mm.

2.4.5 Gaming sessions

Between the two fMRI sessions, participants came seven times to the laboratory to play a video game for one hour. Intervals between subsequent gaming sessions were approximately 24 - 48h, and the second fMRI session was completed at least 24h after the last gaming session. Participants of both groups played a modified version of the game Grand Theft Auto V. In the violent game group, participants controlled a male character equipped with a close-combat weapon, and were tasked to kill as many other characters as possible. Killing was graphically violent, as hitting a character was

accompanied by the splattering of blood, realistic animations of injury, and screams. In the control game group, participants played a version of the game in which all violence was removed. The player character had no weapon, and could not hurt other characters in any way. They could also not be attacked by other characters, and there was no violence between non-player characters. In this condition, participants were tasked to take photographs of as many other characters as possible. In both groups, participants could also freely explore the world of the game. To incentivize a high number of violent or non-violent acts, each kill or photograph was rewarded with one point. For every two points, participants were paid out +0.01€ at the end of the study.

Due to the lack of other studies implementing a randomized experimental prospective design (except for Kühn et al. (2019), published while data collection was already on-going), there were no benchmarks for the amount and frequency of video game exposure for our study. We chose our regimen (seven one-hourly sessions over two weeks) as we considered this a substantial yet still feasible amount of exposure. Number of sessions, playing time per session, and total playing time were considerably higher than in previous studies reporting VVG effects on empathy (Arriaga et al., 2011; Carnagey et al., 2007; Engelhardt et al., 2011; Hasan et al., 2013).

2.4.6 Data analysis

In this paper, we follow a Bayesian data analysis approach (Keysers et al., 2020), which allows clear assessments of the presence or absence of an effect of VVGs on empathy. Hypothesis tests were performed by means of the Bayes Factor (BF; Kass & Raftery, 1995). The BF represents how much more probable the observed data is under the alternative hypothesis compared to the null hypothesis. A well-established convention is to report a BF > 3 as evidence for the alternative hypothesis, a BF < 1/3 as evidence for the null hypothesis, and a BF in the interval [1/3, 3] as inconclusive evidence for either hypothesis (Kass & Raftery, 1995; Keysers et al., 2020). We formulated informed priors for all models to enable valid BF hypothesis tests (Vanpaemel, 2010). To increase comparability with the results of previous papers, we also report analogous frequentist analyses in Appendix 2.E. We registered the analysis plan of this study at https://osf.io/yx423/.

Behavioral data analysis

To test the effects of VVGs on behavioral measures of empathy for pain, we analyzed the VAS ratings obtained during the empathy-for-pain task with hierarchical Bayesian censored regression models. We used censored regression models to account for the fact that participants could give no ratings lower than 0, or higher than 100. Models were estimated using the R package `brms` (Bürkner, 2017). We modelled fixed effects for the experimental factors Group (non-violent vs. violent gaming, coded as -1 and 1), Time (pre vs. post gaming sessions, coded as -1 and 1), and Intensity (non-painful vs. painful stimulation of the confederate, coded as -1 and 1), as well as all interactions between these factors. Additionally, we modelled per-subject random effects of Time, Intensity, and these factors' interaction term. To further account for variations in how participants used the VAS rating scale, we modelled per-subject error variance terms. For further details about the model specification and prior formulation, see Appendix 2.D.

We used the same kind of model to test possible desensitizing effects of VVGs on emotional reactivity to violent images. Here, we modelled fixed effects for the experimental factors Group (non-violent vs. violent gaming, coded as -1 and 1), Content (neutral vs. violent, coded as -1 and 1), and Context (real vs. game, coded as -1 and 1). Additionally, we modelled per-subject random effects for Content, Context, and their interaction, as well as per-subject error variances.

MRI data preprocessing

Preprocessing and analysis of fMRI data were performed using SPM12 (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB 2017b. Functional images were slice timed and referenced to the middle slice, realigned to the mean image, and unwarped using the acquired field map. The structural image was co-registered to the mean image of the realigned functional images using mutual information maximization, and structural and functional images were normalized to the stereotactic Montreal Neurological Institute (MNI) space. The normalized functional images were smoothed with a Gaussian kernel of 4 mm full-width-at-half-maximum, which is equal to twice the voxel size on every axis. To remove motion-related artifacts, the functional images were then subjected to an independent-component-analysis based algorithm for automatic removal of motion artifacts (Pruim, Mennes, Buitelaar, & Beckmann, 2015; Pruim, Mennes, van Rooij, et al., 2015), implemented using the FMRIB software library (FSL v5.0; http://www.fmrib.ox.ac.uk/fsl).

fMRI analyses: empathy for pain

With regards to empathy, our central interest lay in modulations of AI and ACC activity. To identify the regions in which empathic responses were reliably elicited independently of our experimental manipulation, we first analyzed the data from the first experimental session. We performed general-linear model (GLM) based whole-brain analysis using SPM12. For each participant, the design matrix included regressors for the Cue and Stimulation events, separate for all four combinations of conditions (Self No Pain; Self Pain; Other No Pain; Other Pain). As nuisance regressors, we included regressors for the rating events. We then subjected the beta images of the first-level contrast Other Pain > Other No Pain to a one-sample t-test, and identified the voxels in which this contrast was significant and positive ($p < 0.05$ after family-wise error correction). From this, we obtained a binary mask of significant voxels. We then intersected this mask with anatomical masks taken from the Automated Anatomical Labeling atlas (AAL Tzourio-Mazoyer et al., 2002). For the AI ROI, the binary mask was intersected with the AAL mask of the insula (label IN). For the aMCC ROI, the binary mask was intersected with the AAL masks of the anterior and median cingulate and paracingulate gyri (labels ACIN and MCIN). The aim of this masking procedure was to restrict analyses to those parts of the brain areas that are actually recruited by the task. We believe that this increases the sensitivity of our analyses, as we remove signals from voxels that are also part of these anatomical regions, but not actually recruited by the task.

We analyzed signal changes extracted from our ROIs with Bayesian linear mixed-effects model tailored for fMRI data. Note that the ROIs, which were based on the signal from only the first session, were used to extract signals from both sessions. Custom code for this analysis with the software STAN (Carpenter et al., 2017) can be found at https://osf.io/yx423/. See also Appendix 2.D for more information.

The full model included regressors for the Cue and Stimulation events, as well as nuisance regressors for rating events.

fMRI analyses: emotional desensitization

When testing the effects of VVGs on brain activity during the emotional-reactivity task, our main interest lay in a possible modulation of responses in the amygdala, as well as aMCC and AI. To define the corresponding ROIs, we first identified the brain areas that were reliably activated by violent imagery, independent of the experimental manipulation, using whole-brain GLM analysis. For each participant, the design matrix included regressors for the blocks of picture presentations, separate for all four combinations of conditions (Neutral Real, Neutral Game, Violent Real, Violent Game). As nuisance regressors, we included regressors for the rating events. We then pooled the beta images of the first-level contrast *Violent* > *Neutral* across both groups, and subjected them to a one-sample t-test. From this, we obtained a binary mask of voxels significant at $p < 0.05$ after family-wise error correction. We then intersected this mask with AAL masks to obtain our final ROIs (for AI: label IN; for aMCC: labels ACIN and MCIN; for amygdala: label AMYG). We analyzed signal changes extracted from our ROIs with Bayesian linear mixed-effects model. The full model included regressors for the blocks of picture presentation, as well as nuisance regressors for rating events.

2.4.7 Post-hoc analyses

Sample comparability

Due to our preselection of young adult males with minimal prior VVG exposure, it appeared possible that our sample was drawn from a subpopulation with higher trait empathy than the general population. To test this potential limitation, we compared the trait empathy levels of our sample, as measured by the Questionnaire for Cognitive and Affective Empathy (QCAE; Reniers et al., 2011), to those of a control sample of 18-35 year old males who were not preselected for minimal VVG use. The control sample was taken from the dataset of Borghi et al. (2023), which is freely accessible online (https://osf.io/ujp3e). We chose this open dataset because we deemed it highly comparable to our own sample, having also been drawn from the Austrian population, by researchers of the same university. To test whether our sample exhibited higher trait empathy levels than the control sample, we calculated a one-sided Bayesian t-test for each of the five subdimensions of the QCAE, using the R package `BayesFactor` (R. D. Morey & Rouder, 2024).

Test-Retest reliabilities

Given that our experimental design included measurements of participants' empathic responses in two sessions (once before playing the violent video game or the control game, once after), the test-retest reliability ρ of these two measurements was of interest. In our behavioral data, the empathic response in one session was given by the average difference in ratings for *Pain* trials minus *No Pain* trials in session 1 and 2. Given our estimated hierarchical Bayesian censored regression models, the test-retest reliability of empathic responses can be estimated as

$$\rho = \frac{Cov(b_I - b_{I:S}, b_I + b_{I+S}}{\sqrt{Var(b_I - b_{I:S})Var(b_I + b_{I+S})}},$$

where Cov and Var are the Covariance and Variance, respectively, b_I is the random effect of the factor *Intensity* (Pain vs. No Pain), and $b_{I:S}$ is the random effect of the interaction of factors Intensity and Session. By the bilinearity of the covariance operator, this formula can be written in terms of estimated model parameters as

$$\rho = \frac{\sigma^2_{b_I} - \sigma^2_{b_{I:S}}}{\sqrt{(\sigma^2_{b_I} + \sigma^2_{b_{I:S}} - 2r_{b_I b_{I:S}}\sigma_{b_I b_{I:S}})(\sigma^2_{b_I} + \sigma^2_{b_{I:S}} + 2r_{b_I b_{I:S}}\sigma_{b_I b_{I:S}})}},$$

where $\sigma^2_{b_I}$ and $\sigma^2_{b_{I:S}}$ are the variances of the random effect of *Intensity* and *Intensity:Session*, respectively, and where $r_{b_I b_{I:S}}$ is the correlation between these two random effects.

In our neural data, we defined the empathic response in one session as the average difference in BOLD signal to observing the other in pain vs. observing the other in no pain. Given our estimated hierarchical Bayesian regression model, the test-retest reliability of the neural response was given by the correlation coefficient between the random effect for the regressor *Stimulus Other: Pain – No Pain* in Session 1 and the random effect for the equivalent regressor in Session 2.

Bayesian design analysis

We based our sample size on the results of a power analysis designed for the frequentist inference framework (see section Methods: Power analysis). However, as we ultimately based our inference on Bayes factor tests, the theoretical long-term behavior of these tests, given our sample size and expected effect size, is of interest. Therefore, we conducted a post-hoc Bayes Factor design analysis by means of a Monte Carlo simulation experiment (Schönbrodt & Wagenmakers, 2018). The analysis was performed using the R package `BayesFactor` (R. D. Morey & Rouder, 2024). We simulated data from the scenario in which there was no VVG effect on the outcome variable (H0; Cohen's d = 0), as well as from three scenarios where there was a true VVG effect (H1). Here, we considered three different effect sizes: d = 0.4, which is close to the effect size estimate of Anderson et al. (2010; the exact estimate was d = 0.394); d = 0.3, which is the effect size we used in our power analysis; and d = 0.2, the conventional treshold for small effects. For each scenario/effect size, we randomly generated 10000 datasets of the same size as our real sample (control group = 44 participants; experimental group = 45 participants) and subjected them to Bayes factor hypothesis tests, assessing whether the Bayes factor provided evidence for the alternative hypothesis (BF > 3), for the null hypothesis (BF < 1/3), or inconclusive evidence (1/3 < BF < 3). For the behavioral and neural empathy measures, which were measured in two sessions (once before playing the violent video game or the control game, once after), we used test-retest-reliability estimates that are close to those from the previous section.

Cross-task correlations

We additionally report the empirical correlations between the behavioral and neural measurements of our participants empathic response in the empathy-for-pain task, and their response in the emotional reactivity task. As indicators of participants' behavioral responses, we used their estimated random effects from the Bayesian hierarchical models on their rating data (for Empathy for Pain: factor *Intensity*, i.e. Pain vs. No Pain; for Emotional Reactivity: factor *Context*, i.e. Violent vs. Neutral).

As indicators of participants' neural responses, we used their estimated random effects from the Bayesian models on signals extracted from the regions-of-interest (for Empathy for Pain: regressor *Stimulus Other: Pain – No Pain*; for Emotional Reactivity: regressor *Violent – Neutral*).

Data availability

Behavioral data, fMRI signal timecourses extracted from our regions of interest, task event timings, custom STAN code, and game images used in the emotional reactivity task are accessible at https://osf.io/yx423/. Unthresholded statistical maps are accessible at https://identifiers.org/neurovault.collection:13395. These include statistical maps from the analyses underlying the definition of our regions of interest, as well as the statistical maps from the frequentist analyses presented in Appendix 2.E. Full fMRI datasets from all participants are accessible at https://doi.org/10.5281/zenodo.10057633

Acknowledgments

This work was funded in part by the Vienna Science and Technology Fund (WWTF VRG13-007), a Hjärnfonden (FO2014-0189) grant and a Karolinska Institutet 2015 (2-70/2014-97) grant awarded to PP, and a Knut and Alice Wallenberg Foundation (KAW 2014.0237) grant awarded to AO. We would like to thank Sophia Shea, Leonie Brög, and Johannes Ayrle for assistance during data collection.

Conflict of interest

None declared.

2.5 Appendix

2.A Detailed results of the whole-brain analyses for ROI definition

Here, we present the results of the whole-brain analyses underlying our region-of-interest definition (as explained in the main text, section Methods: fMRI analysis: empathy for pain) in more detail. Table 2.A.1 presents the results of the analysis performed on the data from the empathy-for-pain task (only first session). Table 2.A.2 presents the results of the analysis performed on the data from the emotional reactivity task.

TABLE 2.A.1: Results of whole-brain analyses for ROI definition, empathy-for-pain task. Tested contrast: Other Pain - Other No Pain, data only taken from the first session. We report the first local maximum within each cluster. Effects were tested for significance with a significance threshold of p <0.05, FWE-corrected. We only report clusters larger than 10 voxels.

	MNI Coordinates				
Brain Region	**x**	**y**	**z**	**z-value**	**Voxels**
R Insula	30	22	-14	6.15	27
R Inferior frontal gyrus, orbital part	50	26	-6	6.29	125
R Inferior frontal gyrus, opercular part	50	18	8	5.50	23
L Insula	-34	20	0	6.35	343
L Superior frontal gyrus, medial part	-2	34	34	6.32	86
L Supramarginal gyrus	-58	-56	30	6.26	110
L Inferior parietal lobule	58	-56	40	5.69	12

TABLE 2.A.2: Results of whole-brain analyses for ROI definition, emotional reactivity task. Tested contrast: Violent - Neutral. We report the first local maximum within each cluster. Effects were tested for significance with a significance threshold of p < 0.05, FWE-corrected. We only report clusters larger than 10 voxels.

	MNI Coordinates				
Brain Region	**x**	**y**	**z**	**z-value**	**Voxels**
L/R Thalamus	-6	-28	-6	>10	1405
L/R Inferior occipital lobe	-40	-68	-10	>10	30408
R Amygdala	20	0	-14	5.64	26
L Amygdala	-20	-2	-12	6.09	13
L/R Hypothalamus	4	-4	-10	6.41	27
L Precentral gyrus	-44	4	30	>10	521
R Precentral gyrus	46	8	32	8.21	1229
R Caudate	14	12	10	5.80	33
L Inferior frontal gyrus, triangular part	-42	34	16	7.05	159
R Supramarginal gyrus	66	-22	28	6.11	87
L Midcingulum	0	4	34	6.70	34
R Supplementary motor area	6	14	62	6.54	116
L Cerebellum	-22	-38	-42	7.47	42
R Cerebellum	24	-34	-42	6.97	45

2.B Detailed results of Bayesian ROI analyses

In the following, we present the results reported in section Results: fMRI data in more detail. Table 2.B.1 presents the results of the analysis performed on the ROI data extracted from the empathy-for-pain task. Table 2.B.2 presents the results of the analysis performed on the ROI data extracted from the emotional reactivity task.

TABLE 2.B.1: Posterior parameter estimates and contrasts of models for fMRI signal in the empathy-for-pain task. Dependent variable: fMRI signal extracted from the respective ROI (standardized to unit variance): Fixed effect: terms on gray background describe the mean regression parameter of the respective event averaged across all conditions. Other terms describe the fixed effect of the respective condition on the regression parameter. Factor codings: Group: control game group = -1, violent game group = 1; Session: first session = -1, second session = 1; Intensity: non-painful = -1, painful = 1. β/σ_e: Mean model parameter divided by the mean error standard deviation. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

	aMCC			left AI			right AI		
Fixed effect	β/σ_e	95% CI	BF	β/σ_e	95% CI	BF	β/σ_e	95% CI	BF
Self Cue	0.25	(0.09, 0.41)	34.770	0.16	(0.00, 0.32)	1.925	0.33	(0.20, 0.45)	>100
Group	0.03	(-0.13, 0.20)	0.264	-0.01	(-0.17, 0.16)	0.260	0.03	(-0.10, 0.16)	0.257
Intensity	0.14	(0.03, 0.26)	3.494	0.33	(0.20, 0.45)	>100	0.41	(0.31, 0.51)	>100
Session	0.08	(-0.03, 0.19)	0.582	0.10	(-0.01, 0.21)	1.054	0.05	(-0.06, 0.15)	0.301
Group*Intensity	-0.01	(-0.13, 0.10)	0.201	-0.02	(-0.14, 0.11)	0.209	-0.01	(-0.11, 0.09)	0.191
Group*Session	0.05	(-0.05, 0.16)	0.276	0.00	(-0.11, 0.10)	0.179	-0.03	(-0.13, 0.07)	0.236
Intensity*Session	-0.02	(-0.11, 0.08)	0.182	-0.08	(-0.18, 0.02)	0.577	-0.01	(-0.09, 0.07)	0.150
Group*Intensity*Session	-0.03	(-0.12, 0.07)	0.193	-0.04	(-0.14, 0.06)	0.253	-0.03	(-0.11, 0.05)	0.241
Other Cue	0.41	(0.26, 0.56)	>100	0.06	(-0.11, 0.22)	0.299	0.13	(0.00, 0.26)	1.459
Group	0.11	(-0.04, 0.26)	0.627	0.03	(-0.14, 0.21)	0.292	0.09	(-0.04, 0.22)	0.601
Intensity	0.14	(0.05, 0.24)	8.814	0.19	(0.10, 0.28)	>100	0.20	(0.12, 0.28)	>100
Session	0.00	(-0.11, 0.10)	0.175	0.08	(-0.02, 0.20)	0.576	0.07	(-0.02, 0.16)	0.596
Group*Intensity	0.00	(-0.10, 0.09)	0.161	0.02	(-0.07, 0.11)	0.165	-0.01	(-0.09, 0.07)	0.169
Group*Session	-0.05	(-0.16, 0.05)	0.307	0.02	(-0.08, 0.13)	0.207	0.01	(-0.08, 0.10)	0.183
Intensity*Session	-0.03	(-0.13, 0.06)	0.198	-0.07	(-0.17, 0.02)	0.516	-0.09	(-0.17, -0.02)	2.283
Group*Intensity*Session	-0.05	(-0.14, 0.05)	0.402†	-0.06	(-0.15, 0.03)	0.547†	-0.01	(-0.09, 0.07)	0.190†
Self Stimulation	1.02	(0.86, 1.17)	>100	1.38	(1.22, 1.55)	>100	0.82	(0.69, 0.95)	>100
Group	-0.16	(-0.32, 0.00)	1.690	0.01	(-0.16, 0.18)	0.286	-0.02	(-0.16, 0.12)	0.245
Intensity	0.53	(0.39, 0.68)	>100	0.63	(0.49, 0.78)	>100	0.64	(0.51, 0.77)	>100
Session	-0.04	(-0.16, 0.08)	0.280	-0.10	(-0.22, 0.02)	0.617	-0.02	(-0.12, 0.08)	0.218
Group*Intensity	0.01	(-0.14, 0.16)	0.253	0.00	(-0.15, 0.14)	0.289	0.00	(-0.13, 0.13)	0.264
Group*Session	0.10	(-0.02, 0.23)	0.737	0.09	(-0.03, 0.21)	0.585	0.10	(0.00, 0.19)	1.162
Intensity*Session	-0.04	(-0.14, 0.05)	0.220	-0.14	(-0.24, -0.04)	5.631	-0.06	(-0.14, 0.02)	0.385
Group*Intensity*Session	-0.03	(-0.12, 0.07)	0.191	-0.01	(-0.11, 0.09)	0.184	-0.04	(-0.11, 0.04)	0.234
Other Stimulation	0.33	(0.19, 0.48)	>100	0.28	(0.10, 0.46)	20.821	0.23	(0.07, 0.38)	12.416
Group	0.13	(-0.03, 0.28)	0.852	0.09	(-0.10, 0.27)	0.490	0.07	(-0.08, 0.23)	0.378
Intensity	0.32	(0.23, 0.41)	>100	0.39	(0.30, 0.48)	>100	0.36	(0.27, 0.45)	>100
Session	0.11	(0.00, 0.22)	1.236	0.07	(-0.04, 0.18)	0.455	0.05	(-0.05, 0.14)	0.286
Group*Intensity	0.05	(-0.04, 0.14)	0.290	-0.03	(-0.12, 0.05)	0.191	0.02	(-0.07, 0.12)	0.231
Group*Session	-0.01	(-0.12, 0.10)	0.174	-0.09	(-0.20, 0.02)	0.663	-0.03	(-0.13, 0.07)	0.235
Intensity*Session	-0.14	(-0.22, -0.04)	16.742	-0.11	(-0.19, -0.02)	2.595	-0.14	(-0.22, -0.06)	19.384
Group*Intensity*Session	-0.01	(-0.10, 0.09)	0.176†	0.01	(-0.08, 0.10)	0.143†	-0.04	(-0.12, 0.04)	0.434†
Rating	4.16	(3.82, 4.48)	>100	5.01	(4.73, 5.28)	>100	3.24	(2.96, 3.52)	>100
Group	0.02	(-0.17, 0.20)	0.289	-0.06	(-0.25, 0.12)	0.335	0.05	(-0.09, 0.18)	0.333
Session	-0.16	(-0.50, 0.18)	0.700	0.10	(-0.19, 0.37)	0.475	-0.06	(-0.34, 0.22)	0.464
Group*Session	0.04	(-0.15, 0.23)	0.294	-0.02	(-0.20, 0.16)	0.303	-0.05	(-0.17, 0.09)	0.309

TABLE 2.B.2: Posterior parameter estimates and contrasts of models for fMRI signal in the emotional reactivity task. Dependent variable: fMRI signal extracted from the respective ROI (standardized to unit variance): Fixed effect: terms on gray background describe the mean regression parameter of the respective event averaged across all conditions. Other terms describe the fixed effect of the respective condition on the regression parameter. Factor codings: Group: control game group = -1, violent game group = 1; Content: neutral = -1, violent = 1; Context: real = -1, game = 1. β/σ_e: Mean model parameter divided by the mean error standard deviation. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

	left amygdala			right amygdala		
Fixed effect	β/σ_e	95% CI	BF	β/σ_e	95% CI	BF
Pictures	3.59	(3.21, 3.96)	>100	4.25	(3.91, 4.57)	>100
Group	0.22	(-0.15, 0.57)	0.639	-0.02	(-0.33, 0.30)	0.298
Content	1.13	(0.87, 1.38)	>100	1.06	(0.78, 1.31)	>100
Context	-0.01	(-0.23, 0.22)	0.252	-0.06	(-0.29, 0.17)	0.281
Group*Content	-0.04	(-0.29, 0.22)	0.324†	-0.02	(-0.27, 0.25)	0.338†
Group*Context	-0.16	(-0.38, 0.07)	0.554	-0.25	(-0.48, -0.02)	2.204
Content*Context	-0.06	(-0.23, 0.12)	0.227	0.00	(-0.18, 0.18)	0.205
Group*Content*Context	-0.01	(-0.18, 0.17)	0.205†	0.02	(-0.17, 0.20)	0.163†
Rating	1.02	(0.73, 1.30)	>100	0.97	(0.72, 1.24)	>100
Group	-0.10	(-0.36, 0.19)	0.315	-0.05	(-0.30, 0.21)	0.255

2.C Pictures used in the emotional reactivity task

We conducted a pilot study to match pictures taken from the video game to pictures taken from the International Affective Picture System (IAPS; Lang et al., 2005). We preselected 33 IAPS pictures of neutral content (people with neutral facial expressions, objects) and 34 IAPS pictures of violent content (dead bodies, mutilations, fights, weapons), as well as 33 game pictures of neutral content and 46 games pictures of violent content. In an online survey, thirty-one participants (16 female, 15 male) rated these pictures in terms of valence and arousal, using the 9-point self-assessment manikin scale (Bradley & Lang, 1994).

We calculated the mean values of valence and arousal across participants per picture, and used these scores, as well as the individual pictures' content, to select ten pictures per condition as stimuli for the emotional reactivity task. The scores for the final selection of pictures are listed in Table 2.C.1. Note that after matching, there were still systematic differences between game and real pictues: violent real pictures were generally rated higher in arousal, and lower in valence, than violent game pictures. This is due to the fact that a matching purely on valence and arousal scores would have led to sets of pictures with highly different contents (i.e., real pictures showing mostly fights and threats without blood, and game pictures showing mostly dead bodies and highly violent attacks). However, we deemed it important that real pictures and game pictures were also as similar in content as possible. Moreover, we believe that a difference in valence and arousal between real pictures and game pictures is only a minor issue for the experimental design. Our main research question does not concern differences in behavioral and neural responses to real vs. game pictures, but how these responses differ between participants who played a highly violent video game and participants who played a non-violent video game. The game pictures are available at https://osf.io/yx423/ .

TABLE 2.C.1: Pictures used in the emotional reactivity paradigm. Picture ID: IAPS ID numbers for pictures of the Neutral Real and Violent Real conditions; arbitrary internal ID numbers for pictures of the Neutral Game and Violent Game conditions.

Condition	Picture ID	Content	Valence	Arousal
Neutral Real	2038	Woman reading alone	5.60	2.17
	2200	Man	5.52	1.71
	2210	Man	5.10	2.00
	2215	Man	5.30	1.63
	2383	Woman on the phone	5.33	1.30
	2393	Two workers	5.37	1.37
	2440	Woman	5.48	1.42
	2495	Man	5.27	1.87
	2570	Man	5.50	1.57
	2749	Man smoking	5.33	1.90
Violent Real	3010	Dead body	1.45	6.81
	3015	Dead body	1.32	7.16
	3016	Dead body	1.87	6.20
	3060	Mutilation	1.42	7.00
	3120	Mutilation	1.63	6.27
	6530	Man hits woman	2.81	4.55
	6550	Man threatens woman with knife	2.06	5.7
	6560	Man threatens woman with gun	1.87	6.13
	6561	Man hits woman	3.29	4.23
	6571	Man threatens man with gun	3.23	4.06
Neutral Game	ng1	Woman	5.40	1.67
	ng2	Three women smoking	5.32	1.84
	ng3	Man	5.13	1.68
	ng4	Woman	5.40	1.80
	ng5	Woman	5.39	1.35
	ng6	Man	5.23	1.42
	ng7	Woman	5.61	1.81
	ng8	Man	5.45	1.48
	ng9	Man	5.65	1.48
	ng10	Two workers	5.52	1.52
Violent Game	vg1	Man attacks man with chainsaw	2.13	5.90
	vg2	Mutilation	2.23	5.65
	vg3	Dead body	2.43	5.33
	vg4	Dead body	2.52	5.16
	vg5	Dead body	2.52	4.71
	vg6	Man shoots man in the head	2.63	5.00
	vg7	Man shoots man in the head	2.58	5.29
	vg8	Man chokes man	3.13	4.37
	vg9	Man shoots man in the head	3.06	4.52
	vg10	Man threatens man with gun	3.61	3.74

2.D Frequentist analyses

Behavioral data

As equivalent frequentist analyses of our behavioral data, we fitted linear mixed effects models to the collected ratings. All Models were estimated using the R package *lmerTest* (Kuznetsova et al., 2017). P-values were derived using the Satterthwaite approximation of degrees of freedom. We used the conventional significance level of $\alpha = 0.05$, and one-sided testing for directional hypotheses.

Empathy for pain

To analyze the painfulness and unpleasantness ratings collected during the empathy-for-pain paradigm, we modelled fixed effects for the experimental factors *Group* (non-violent vs. violent gaming, coded as -1 and 1), *Time* (pre vs. post gaming sessions, coded as -1 and 1), and *Intensity* (non-painful vs. painful stimulation of the confederate, coded as -1 and 1), as well as all interactions between these factors. Additionally, we modelled per-subject random effects of *Time*, *Intensity*, and these factors' interaction term. Table 2.D.1 presents the results of these analyses. For both ratings, we observed a non-significant *Group*Session*Intensity* interaction (for painfulness: onesided p-value = .080; for unpleasantness: onesided p-value = .381), implying no evidence for a VVG effect on behavioral correlates of empathy for pain.

TABLE 2.D.1: Linear mixed effects models for ratings in the empathy-for-pain task. Dependent variable: empathy ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Session: first session = -1, second session = 1; Intensity: non-painful = -1, painful = 1.

Fixed effect	β	**SE**	**df**	**t**	**p**
Painfulness ratings					
Group	0.03	0.74	87.7	0.04	.967
Session	0.16	0.43	85.9	0.38	.706
Intensity	25.49	0.98	85.7	25.93	<.001
Group*Session	-0.68	0.43	85.9	-1.56	.122
Group*Intensity	-0.92	0.98	85.7	-0.96	.340
Session*Intensity	-0.23	0.47	82.0	-0.50	.621
Group*Session*Intensity	-0.67	0.47	82.0	-1.42	.160
Unpleasantness ratings					
Group	-1.06	1.46	86.36	-0.73	.469
Session	-0.81	0.55	82.36	-1.47	.145
Intensity	14.97	1.11	86.68	13.51	<.001
Group*Session	-1.04	0.55	82.36	-1.89	.063
Group*Intensity	-0.43	1.11	86.68	-0.39	.696
Session*Intensity	-0.99	0.43	84.78	-2.31	.023
Group*Session*Intensity	-0.13	0.43	84.78	-0.30	.762

Emotional reactivity

To analyze the unpleasantness ratings collected during the emotional reactivity paradigm, we modelled fixed effects for the experimental factors *Group* (non-violent vs. violent gaming, coded as -1 and 1), *Content* (neutral vs. violent, coded as -1 and 1), and *Context* (real vs. game, coded as -1 and 1). Additionally, we modelled per-subject random effects for *Content*, *Context*, and their interaction. Table 2.D.2 presents the

TABLE 2.D.2: Linear mixed effects models for unpleasantness ratings in the emotional reactivity task. Dependent variable: unpleasantness ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Content: neutral = -1, violent = 1; Context: real = -1, game = 1.

Fixed effect	β	**SE**	**df**	**t**	**p**
Group	0.96	1.48	87	0.65	.519
Content	29.17	1.50	87	19.40	<.001
Context	-5.03	0.71	87	-7.10	<.001
Group*Content	1.49	1.50	87	0.99	.325
Group*Context	-0.58	0.71	87	-0.82	.415
Content*Context	-4.80	0.67	87	-7.18	<.001
Group*Content*Context	-0.03	0.67	87	-0.05	.962

results of these analyses. We observed no significant Group*Content interaction (one-sided p-value = .163) or Group*Content*Context interaction (one-sided p-value = .481), implying no evidence for a VVG effect on behavioral correlates on emotional responses to violent images.

FMRI data

We performed general-linear model (GLM) based whole-brain analysis using SPM12 (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm), implemented in MATLAB 2017b. Parameter estimates were estimated on the first level, and the contrasts of interest were then subjected to two-sample t-tests on the second level. To increase power to detect effects in our a-priori defined regions of interests (ROI), we used small-volume correction, using the same ROIs as for the bayesian analyses in the main text. We tested for voxels that survived family-wise error correction, $p < .05$. To give a more complete picture, we also tested for voxels inside the ROIs that survived the more lenient thresholds of $p < .001$ uncorrected and $p < .05$ uncorrected.

Empathy for pain

For each participant and session, the first level design matrix included regressors for the Cue and Stimulation events, separate for all four combinations of conditions (Self No Pain; Self Pain; Other No Pain; Other Pain). As nuisance regressors, we included regressors for the rating events. We then subjected the beta images of the first-level contrast $[\textit{Other Pain - Other No Pain}]_{\text{Session 2}}$ - $[\textit{Other Pain - Other No Pain}]_{\text{Session 1}}$ to a two-sample t-test, and identified the voxels in which the contrast *Control Group* > *Experimental Group* was significant and positive.

Using family-wise error correction, we found no significant clusters in any of the three ROIs (AMCC, left AI, right AI). There were also no voxels surviving the uncorrected threshold of $p < .001$. In left AI, one voxel out of 343 survided the uncorrected threshold of $p < 0.05$.

Other whole-brain results may be investigated using the T-map provided online (https://identifiers.org/neurovault.collection:13395).

Emotional reactivity

For each participant, the design matrix included regressors for the blocks of picture presentations, separate for all four combinations of conditions (Neutral Real, Neutral

Game, Violent Real, Violent Game). As nuisance regressors, we included regressors for the rating events. We then subjected the beta images of the first-level contrasts of interest to a two-sample t-test, and identified the voxels in which the contrast *Control Group* > *Experimental Group* was significant and positive.

For the interaction *Group*Content* (testing whether participants in the violent game group had decreased responses to violent images in general), the contrast of interest was [*Violent Real* + *Violent Game*]/2 - [*Neutral Real* + *Neutral Game*]/2. Using family-wise error correction, we found no significant clusters in any of the two ROIs (left amygdala, right amygdala). There were also no voxels surviving the uncorrected threshold of $p < .001$. In the left amygdala, 3 voxels out of 220 survided the uncorrected threshold of $p < 0.05$, and in the right amygdala, 10 voxels out of 248 survived this more lenient threshold.

For the interaction *Group*Content*Context* (testing whether participants in the violent game group had decreased responses to specifically violent game images), the contrast of interest was [*Violent Game - Neutral Game*] - [*Violent Real - Neutral Real*]. Using family-wise error correction, we found no significant clusters in any of the two ROIs (left amygdala, right amygdala). There were also no voxels surviving the uncorrected threshold of $p < .001$. In the left amygdala, 15 voxels out of 220 survided the uncorrected threshold of $p < 0.05$, and in the right amygdala, 49 voxels out of 248 survived this more lenient threshold.

Other whole-brain results may be investigated using the T-maps provided online (https://identifiers.org/neurovault.collection:13395).

2.E Bayesian hierarchical models

Behavioral data analysis

We fitted hierarchical censored regression models to the rating data from the empathy-for-pain task and the emotional reactivity task. Ratings were collected using a 100-step visual analog scale (VAS), and could thus lie in the range [0,100]. For numerical reasons, we first linearly transformed ratings to the range [-3,3]. In the following, let i index participants, and t index trials. The censored regression model relates y_{it}, the rating given by participant i in trial t, to a latent response variate y^*_{it} by the function

$$y_{it} = \begin{cases} -3, & y^*_{it} \leq -3 \\ y^*_{it}, & -3 < y^*_{it} < 3 \\ 3, & 3 \leq y^*_{it} \end{cases}.$$

For y^*_i, the vector of latent responses of participant i, we formulate the linear model

$$y^*_i = \mu + X_i\beta + Z_i b_i + \varepsilon_i,$$

where μ is the grand mean parameter, X_i and Z_i are the i-th participant's design matrices associated with fixed effects and random effects, respectively, β is the vector of fixed effects, b_i is the vector of random effects of participant i, and ε_i is the vector of error terms. Further, we assume

$$\begin{aligned} \varepsilon_{it} &\sim \mathcal{N}\left(0, \sigma_{\varepsilon_i}\right) \\ \sigma_{\varepsilon_i} &\sim \text{Lognormal}\left(\mu_{\sigma_\varepsilon}, \sigma_{\sigma_\varepsilon}\right), \end{aligned}$$

where σ_{ε_i} is the residual error variance associated with participant i, and μ_{σ_ε} and $\sigma_{\sigma_\varepsilon}$ are the hyperparameters of the Lognormal distribution of residual error variances. For these hyperparameters, we formulate the priors

$$\begin{aligned} \mu_{\sigma_\varepsilon} &\sim \mathcal{N}\left(0, \frac{1}{2}\right) \\ \sigma_{\sigma_\varepsilon} &\sim \text{Gamma}\left(4 \cdot \log(2), 4\right), \end{aligned}$$

which put the majority of their mass on sensible values. For the vector of random effects , we assume

$$\begin{aligned} b_i &\sim \mathcal{N}_{k_b}(0, D \cdot R \cdot D), \\ D &= \text{diag}\left(\sigma_{b_1}, ..., \sigma_{b_{kb}}\right) \end{aligned}$$

where k_b is the number of random effects, D is the diagonal matrix with the random effect standard deviations $\sigma_{b_1}, ..., \sigma_{b_{kb}}$ on the diagonal, and R is the correlation matrix of random effects. We further formulate the weakly informative priors

$$\begin{aligned} \sigma_{b_1}, ..., \sigma_{b_{kb}} &\sim \text{Halfnormal}(0, 1), \\ R &\sim \text{LKJ}(2). \end{aligned}$$

Lastly, we formulate the following prior on the fixed effects,

$$\begin{aligned}\mu &\sim \mathcal{N}(0,3),\\ \frac{\beta}{\bar{\sigma}_\varepsilon} &\sim \mathcal{N}_{k_\beta}\left(0, \frac{1}{2}\cdot I_{k_\beta}\right),\\ \bar{\sigma}_\varepsilon &= \exp\left(\mu_{\sigma_\varepsilon} + \frac{\sigma^2_{\sigma_\varepsilon}}{2}\right)\end{aligned}$$

where k_β is the number of fixed effects, and $\bar{\sigma}_\varepsilon$ is the theoretical mean of error standard deviations across participants. Putting the prior on the ratio $\frac{\beta}{\bar{\sigma}_\varepsilon}$ instead of β allows us to formulate an appropriately informed prior without prior knowledge of the average variance of the error term. We use the scaling factor $\frac{1}{2}$ to represent our prior assumption that fixed effects are unlikely to be much larger (in absolute value) than the average error standard deviation.

FMRI data analysis

We fitted hierarchical regression models to the BOLD response data extracted from the regions-of-interest. To account for the autocorrelation that is to be expected in fMRI data, we assumed that the residual error terms within a run could be described by an autoregressive process of order 1 (AR(1)).

To facilitate interpretation and formulation of priors, we directly parameterized the model in terms of contrasts of regression weights. For each subject i and session j, we first build the raw design matrix X^R_{ij} as is done in established software such as SPM (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm). Shortly, for each separate type of event, we created a regressor representing the expected BOLD signal induced by the event by convolving a boxcar function of appropriate onset and length with the canonical hemodynamic response function. Then we constructed the design matrix in terms of contrasts X^C_{ij} by right-multiplying X^R_{ij} with a matrix C that encoded the contrasts of interest. For example, consider the raw design matrix X^R_{ij} of the empathy-for-pain paradigm with the following mapping between column numbers and events:

$1 \mapsto$ Cue Self: No Pain
$2 \mapsto$ Cue Self: Pain
$3 \mapsto$ Cue Other: No Pain
$4 \mapsto$ Cue Other: Pain
$5 \mapsto$ Stimulus Self: No Pain
$6 \mapsto$ Stimulus Self: Pain
$7 \mapsto$ Stimulus Other: No Pain
$8 \mapsto$ Stimulus Other: Pain
$9 \mapsto$ Rating.

Then

$$X_{ij}^R C = X_{ij}^R \begin{pmatrix} 1 & -1 & & & & & & & \\ 1 & 1 & & & & & & & \\ & & 1 & -1 & & & & & \\ & & 1 & 1 & & & & & \\ & & & & 1 & -1 & & & \\ & & & & 1 & 1 & & & \\ & & & & & & 1 & -1 & \\ & & & & & & 1 & 1 & \\ & & & & & & & & 1 \end{pmatrix}$$

results in a matrix X_{ij}^C with the column-to-contrast mapping

$$\begin{aligned} 1 &\mapsto \text{Cue Self: Mean} \\ 2 &\mapsto \text{Cue Self: Pain - No Pain} \\ 3 &\mapsto \text{Cue Other: Mean} \\ 4 &\mapsto \text{Cue Other: Pain - No Pain} \\ 5 &\mapsto \text{Stimulus Self: Mean} \\ 6 &\mapsto \text{Stimulus Self: Pain - No Pain} \\ 7 &\mapsto \text{Stimulus Other: Mean} \\ 8 &\mapsto \text{Stimulus Other: Pain - No Pain} \\ 9 &\mapsto \text{Rating.} \end{aligned}$$

Finally, for tasks with two sessions (i.e. the empathy-for-pain task), main effects and interactions with the session factor were represented by constructing the final first-level design matrix as the block matrix

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} = \begin{pmatrix} X_{i1}^C & -X_{i1}^C \\ X_{i2}^C & X_{i2}^C \end{pmatrix}.$$

With this construction, the first half of columns of X_i correspond to the effects of events marginal to the session factor, while the second half of columns correspond to the interactions between events and the session factor.

For describing the hierarchical model, we denote the sequence of extracted signals of participant i in session j as y_{ij}. Further, we let t index the timepoints within each such sequence, such that $y_{ij} = \left(y_{ij1}, y_{ij2}, ..., y_{ijt}, ...\right)'$. To account for low-frequency changes of signal of no interest (e.g., scanner drift), each y_{ij} was first filtered with a high-pass filter of period 128. As differences in grand mean between participants and sessions were not of interest, each sequence y_{ij} was mean centered to have mean 0. Further, to change the arbitrary scale of BOLD response signals to a known scale, each y_{ij} was scaled to have variance 1. The same operations were also performed on each design matrix X_{ij}.

For each y_{ij}, we formulate the linear model

$$y_{ij} = X_{ij} \cdot (\beta + G_i\gamma + b_i) + u_{ij},$$

where β is the vector of fixed effects of contrasts, G_i is a variable that takes the value -1 when participant i is in the control group, and 1 when participant i is in the experimental group, γ is the vector of fixed effects of the factor group, b_i is the vector

of random effects of participant i, and u_{ij} is the vector of error terms of participant i in session j. We assume that u_{ij} follows an AR(1) process, thus

$$\begin{aligned} u_{ijt} &= \varepsilon_{ijt} + \varphi_{ij}\varepsilon_{ij(t-1)}, \\ \varepsilon_{ijt} &\sim \mathcal{N}\left(0, \sigma_{\varepsilon ij}\right), \end{aligned}$$

where φ_{ij} is the autoregression parameter of person i in session j, ε_{ijt} are independently and identically distributed impulses, and $\sigma_{\varepsilon ij}$ is the standard deviation of these impulses for participant i and session j. Following the hierarchical modeling approach, we assume that φ_{ij} and $\sigma_{\varepsilon ij}$ are themselves drawn from a higher-order distribution, for whose hyperparameters we formulate weak priors:

$$\begin{aligned} \frac{\varphi_{ij}+1}{2} &\sim \text{Logitnormal}\left(\mu_\varphi, \sigma_\varphi\right), \\ \mu_\varphi &\sim \mathcal{N}(0,1), \\ \sigma_\varphi &\sim \text{Halfnormal}(0,1), \end{aligned}$$

and

$$\begin{aligned} \sigma_{\varepsilon ij} &\sim \text{Lognormal}\left(\mu_{\sigma_\varepsilon}, \sigma_{\sigma_\varepsilon}\right), \\ \mu_{\sigma_\varepsilon} &\sim \mathcal{N}(-1/2, 1/2), \\ \sigma_{\sigma_\varepsilon} &\sim \text{Gamma}\left(4 \cdot log(2), 4\right). \end{aligned}$$

For the vector of random effects , we assume

$$\begin{aligned} b_i &\sim \mathcal{N}_{k_b}(0, D \cdot R \cdot D), \\ D &= \text{diag}\left(\sigma_{b_1}, ..., \sigma_{b_{kb}}\right) \end{aligned}$$

where k_b is the number of random effects, D is the diagonal matrix with the random effect standard deviations $\sigma_{b_1}, ..., \sigma_{b_{kb}}$ on the diagonal, and R is the correlation matrix of random effects. We further formulate the priors

$$\begin{aligned} R &\sim \text{LKJ}(2), \\ \sigma_{b_1}, ..., \sigma_{b_{kb}} &\sim \text{Halfnormal}\left(0, \frac{\bar{\sigma}_\varepsilon}{2}\right), \\ \bar{\sigma}_\varepsilon &= \exp\left(\mu_{\sigma_\varepsilon} + \frac{\sigma^2_{\sigma_\varepsilon}}{2}\right), \end{aligned}$$

where $\bar{\sigma}_\varepsilon$ is the theoretical mean of error standard deviations across participants. This reflects the assumption that random effects will not be much larger than the mean error standard deviation. Lastly, we formulate the following priors on the fixed effects,

$$\begin{aligned} \frac{\beta}{\bar{\sigma}_\varepsilon} &\sim \mathcal{N}_{k_\beta}\left(0, \frac{1}{10} \cdot I_{k_\beta}\right), \\ \frac{\gamma}{\bar{\sigma}_\varepsilon} &\sim \mathcal{N}_{k_\gamma}\left(0, \frac{1}{10} \cdot I_{k_\gamma}\right), \end{aligned}$$

where k_β and k_γ are the number of fixed effects. Putting the prior on the ratio $\frac{\beta}{\bar{\sigma}_\varepsilon}$ instead of β allows us to formulate an appropriately informed prior without prior knowledge of the average variance of the error term. We use the scaling factor $\frac{1}{10}$ to

represent our prior assumption that fixed effects are likely to be much smaller (in absolute value) than the average error standard deviation, due to the high amount of noise in fMRI signal. However, as this prior formulation might still have been too vague for proper hypothesis testing via the Bayes Factor, we additionally informed the prior with a fraction of $\frac{2}{n}$ of the likelihood of the data (where n is the sample size), therefore calculating fractional Bayes factors (O'Hagan, 1995).

2.F Covariate analyses

As described in our registration, we additionally performed analyses to investigate the role of trait neuroticism and executive control on the possible violent video game (VVG) effect on empathy. As a measure of trait neuroticism, we used the Neuroticism scale of the German version of the NEO-FFI (Borkenau & Ostendorf, 1993). Due to technical issues, the neuroticism measure could not be obtained from seven participants, leaving a sample size of N = 82 for these analyses. As a measure of executive control, we used stop-signal reaction time (SSRT), measured with the software STOP-IT (Verbruggen et al., 2008). Due to technical issues, valid SSRT measures of 8 participants were missing. Additionally, we removed SSRT measures of 3 participants who inhibited responses in significantly more or less than 50% of times (see Verbruggen et al., 2008, for an explanation of this criterion). In total, the sample size for analyses involving SSRT was N = 78.

We added the fixed effects of Neuroticism/SSRT, as well as its interactions with other factors, to the models described in section *Methods: Behavioral data analysis*. Results are shown in 2.F.1 for covariate Neuroticism, and in 2.F.2 for covariate SSRT. The Bayes factor of the tests of the interaction *Neuroticism*Group*Intensity*Session* was 0.265 for painfulness ratings, and 0.466 for unpleasantness ratings. The Bayes factor of the tests of the interaction *SSRT*Group*Intensity*Session* was 0.128 for painfulness ratings, and 0.021 for unpleasantness ratings. This indicates that behavioral VVG effects could also not be observed in participants with high levels of trait neuroticism resp. high levels of SSRT.

To analyze neural responses, we added the fixed effects of Neuroticism/SSRT, as well as its interactions with other factors, to the models described in section *Methods: fMRI analyses: empathy for pain*. For *Other Cue* events, the Bayes factors of the tests of the interaction *Neuroticism*Group*Intensity*Session* were $BF_{aMCC} = 1.517$; $BF_{left\ AI} = 0.901$; $BF_{right\ AI} = 0.703$. For *Other Stimulation* events, the Bayes factors of the tests of the same interaction were $BF_{aMCC} = 0.348$; $BF_{left\ AI} = 0.226$; $BF_{right\ AI} = 0.209$. Thus, our data give mixed levels of evidence for the absence of a modulation of VVG effects through trait neuroticism. For *Other Cue* events, the data is inconclusive: we cannot confidently say that there is indeed no modulation of the VVG effect on brain activity during cues that indicate whether or not the other person will receive a painful stimulus. For *Other Stimulation* events, we obtain moderate evidence for the absence of such a modulation: we can, with some confidence, say that participants with high neuroticism were not more susceptible to VVG effects on brain activity while the other person received painful stimulation.

For *Other Cue* events, the Bayes factors of the tests of the interaction *SSRT * Group * Intensity * Session* were $BF_{aMCC} = 0.087$; $BF_{left\ AI} = 0.126$; $BF_{right\ AI} = 0.329$. For *Other Stimulation* events, the Bayes factors of the tests of the same interaction were $BF_{aMCC} = 0.936$; $BF_{left\ AI} = 0.316$; $BF_{right\ AI} = 0.551$. Thus, our data give mixed levels of evidence for the absence of a modulation of VVG effects through executive control.

For *Other Cue* events, we obtain moderate to substantial evidence for the absence of such a modulation: we can, with some confidence, say that participants with low executive control were not more susceptible to VVG effects on brain activity during cues that indicate whether or not the other person will receive a painful stimulus. For *Other Cue* events, the data is inconclusive: we cannot confidently say that there is indeed no modulation of the VVG effect on brain activity while the other person received painful stimulation.

TABLE 2.F.1: Posterior parameter means of models for ratings in the empathy-for-pain task, including the trait covariate Neuroticism. Dependent variable: empathy ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Session: first session = -1, second session = 1; Intensity: non-painful = -1, painful = 1. The variable Neuroticism was scaled to mean 0 and standard deviation 1. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

Fixed effect	β	**95% Credible Interval**		**Bayes Factor**
Painfulness ratings				
Group	0.622	-1.147	2.385	0.138
Neuroticism	-0.624	-2.401	1.135	0.140
Session	0.518	-0.240	1.270	0.105
Intensity	27.208	24.593	29.772	>100
Group*Neuroticism	-1.767	-3.597	0.024	0.541
Group*Session	-0.506	-1.278	0.239	0.100
Neuroticism*Session	-0.691	-1.411	0.009	0.225
Group*Intensity	-1.253	-3.788	1.276	0.255
Neuroticism*Intensity	0.537	-2.022	3.069	0.174
Session*Intensity	-0.374	-1.521	0.791	0.082
Group*Neuroticism*Session	-0.155	-0.898	0.578	0.046
Group*Neuroticism*Intensity	-0.383	-2.999	2.205	0.158
Group*Intensity*Session	-0.599	-1.743	0.506	0.112
Neuroticism*Intensity*Session	0.628	-0.536	1.826	0.113
Group*Neuroticism*Intensity*Session	-0.717	-1.865	0.473	0.265†
Unpleasantness ratings				
Group	-0.562	-4.120	3.058	0.202
Neuroticism	3.430	-0.141	7.296	1.007
Session	-0.258	-1.534	1.131	0.081
Intensity	17.004	14.246	19.772	>100
Group*Neuroticism	-2.029	-5.759	1.932	0.390
Group*Session	-0.734	-2.060	0.589	0.127
Neuroticism*Session	-1.682	-3.021	-0.377	1.697
Group*Intensity	-0.875	-3.536	1.770	0.181
Neuroticism*Intensity	1.977	-0.917	4.697	0.390
Session*Intensity	-1.192	-2.124	-0.269	1.236
Group*Neuroticism*Session	-0.904	-2.301	0.443	0.168
Group*Neuroticism*Intensity	0.496	-2.258	3.253	0.167
Group*Intensity*Session	-0.246	-1.161	0.696	0.060
Neuroticism*Intensity*Session	-0.464	-1.402	0.486	0.081
Group*Neuroticism*Intensity*Session	-0.850	-1.796	0.092	0.466†

TABLE 2.F.2: Posterior parameter means of models for ratings in the empathy-for-pain task, including the trait covariate SSRT (signal-stop reaction time). Dependent variable: empathy ratings (visual analog scale, range: 0 - 100). Factor codings: Group: control game group = -1, violent game group = 1; Session: first session = -1, second session = 1; Intensity: non-painful = -1, painful = 1. The variable SSRT was scaled to mean 0 and standard deviation 1. Bayes Factors were derived from comparing a model where the respective parameter was unrestricted to a model where it was restricted to zero. † These Bayes factors were derived from comparing a model where the parameter was restricted to be negative to a model where it was restricted to zero (one-sided hypothesis test).

Fixed effect	β	**95% Credible Interval**		**Bayes Factor**
Painfulness ratings				
Group	0.990	-1.016	3.014	0.173
SSRT	0.060	-2.183	2.346	0.128
Session	0.529	-0.374	1.431	0.101
Intensity	27.712	25.107	30.359	>100
Group*SSRT	-0.413	-2.724	1.953	0.139
Group*Session	-0.539	-1.517	0.363	0.097
SSRT*Session	-0.290	-1.389	0.794	0.068
Group*Intensity	-1.590	-4.051	0.933	0.294
SSRT*Intensity	0.291	-2.671	3.222	0.165
Session*Intensity	0.051	-1.035	1.145	0.063
Group*SSRT*Session	0.587	-0.440	1.611	0.105
Group*SSRT*Intensity	2.030	-0.898	4.986	0.402
Group*Intensity*Session	-0.830	-1.944	0.230	0.173
SSRT*Intensity*Session	1.138	-0.170	2.400	0.333
Group*SSRT*Intensity*Session	-0.409	-1.710	0.861	0.128†
Unpleasantness ratings				
Group	-0.080	-4.225	3.970	0.214
SSRT	0.661	-3.864	5.040	0.262
Session	-0.476	-1.878	1.018	0.103
Intensity	17.149	14.265	19.969	>100
Group*SSRT	-3.128	-7.644	1.276	0.613
Group*Session	-0.817	-2.328	0.692	0.151
SSRT*Session	0.688	-0.962	2.386	0.127
Group*Intensity	-1.531	-4.313	1.325	0.259
SSRT*Intensity	1.193	-1.907	4.284	0.216
Session*Intensity	-1.089	-2.014	-0.211	0.822
Group*SSRT*Session	1.253	-0.465	2.938	0.269
Group*SSRT*Intensity	1.527	-1.569	4.545	0.277
Group*Intensity*Session	-0.092	-1.015	0.834	0.051
SSRT*Intensity*Session	-0.137	-1.136	0.900	0.055
Group*SSRT*Intensity*Session	0.999	-0.040	2.031	0.021†

Chapter 3

With Low Power comes Low Credibility? Towards a Principled Critique of Results from Underpowered Tests

Abstract

Researchers should be motivated to adequately power statistical tests, as tests with low power have a low probability of detecting true effects. However, it is also often claimed that significant results obtained by underpowered tests are less likely to reflect a true effect. Here, we critically discuss this "low power - low credibility" (LPLC) critique from both a frequentist and Bayesian perspective. Although the LPLC critique is first and foremost a critique of frequentist tests, it is itself not consistent with frequentist theory. In particular, it demands that we have some information on the probability that a hypothesis is true before we test it. However, such prior probabilities are dismissed as meaningless in frequentist inference, and we demonstrate that they cannot be meaningfully introduced into frequentist thinking. Even when adopting a Bayesian perspective, however, significant results from tests with low power can provide a non-negligible amount of support for the tested hypothesis. We conclude that, even though low power reduces the chances to obtain significant findings, there is little justification to dismiss already obtained significant findings on the basis of low power alone. However, concerns about low power might often reflect suspicions that findings were produced by questionable research practices. If this is the case, we suggest that communicating these issues transparently rather than using low power as a proxy concern will be more appropriate. We conclude by providing suggestions on how results from tests with low power can critiqued for the correct reasons and in a constructive manner.

3.1 Introduction

Psychological science is in a period of reform. In the wake of the replication crisis, empirical researchers and methodologists have raised awareness about questionable research practices (QRPs), pointing out that many analytical habits decrease the credibility of scientific output (Bakker et al., 2012; Open Science Collaboration, 2015). Most reform efforts address the (mis)use of frequentist inference, especially in the form of significance testing and p-values (McShane et al., 2019; Wicherts et al., 2016).

This chapter is published as **Lengersdorff, L. L.** & Lamm, C. (2025). With Low Power comes Low Credibility? Towards a Principled Critique of Results from Underpowered Tests. *Advances in Methods and Practices in Psychological Science*, *8*(1). https://doi.org/10.1177/25152459241296397

One of the biggest concerns raised is that many of the conducted statistical tests suffer from low power (Button et al., 2013; Ioannidis, 2005; Lindsay, 2015). Per definition, a statistical test with low power has a low probability of producing a significant result, even when the tested hypothesis is true. On these grounds alone, every researcher should be discouraged from conducting studies with sample sizes that are too small to detect plausible effect sizes, as they will likely result in a waste of resources and time. However, it now seems a widely-held belief that underpowered tests are problematic for another reason : Not only does low power decrease the probability of a significant result when the hypothesis is true, but (so goes the notion) it also makes the significant results that *are* obtained less credible, because they have a lower probability of reflecting a true finding. This idea is prominently voiced in Ioannidis's influential paper 2005, as one of the reasons "why most published research findings are false" (p. 696), and cited and further expanded in numerous other papers (e.g., Button et al., 2013; Colquhoun, 2014)(but see Wagenmakers et al., 2015, for a critical evaluation). We will from now on refer to this as the "Low Power - Low Credibility" (LPLC) critique.

In this paper, we sketch out the development of the LPLC critique, and discuss the assumptions one needs to accept to use it, from both a frequentist and Bayesian point of view. Importantly, we also debate whether or not results from tests with low power are particularly susceptible to being invalidated by the use of QRPs, and evaluate the LPLC critique in the presence of publication bias. We expect that researchers with a general interest in statistical inference will benefit from reading this paper (if only by the opportunity to reflect on why they disagree with our points). We would like to achieve two things: First, to make scientists reflect, question, and openly communicate their own assumptions when they use the LPLC critique; and second, to empower scientists to demand the same level of reflection and open communication from peers who use the LPLC critique when assessing their work, e.g. during peer review. At the end of this paper, we provide recommendations on how this reflection and communication could be achieved in practice.

We would like to explicitly emphasize here that only an ill-informed reading of our arguments could be used as an "excuse" for deliberately conducting studies with suboptimal sample sizes. After all, tests that have low power for even the largest plausible effect size are very likely to produce non-significant results. We expect that, in the presence of the reflection and open communication we advocate, such studies will be justifiably assessed as irrelevant or inefficient to achieve scientific progress.

3.2 What does "underpowered" mean?

The recent debate about the prevalence of "underpowered" tests can create the impression that power is an objective property of a test. Indeed, in our view, calling a test or study "underpowered" is often used as a shorthand to say that the sample size is too small relative to some explicit or implicit standard. However, one test with one given sample size does not have "one" power, but many different levels of power for different possible *effect sizes* of interest (R. Morey & Lakens, 2016). However, what effect sizes are "interesting" or "plausible" must ultimately be decided by scientists and their peers, and it is through this decision that a substantial amount of subjective consideration enters the assessment of a test's power. Thus, a fair and transparent critique must include a statement for which effect sizes the test's power was too low. As Morey 2019 put it succinctly, " 'This sample size feels small' is not good enough"

(section *Critiquing power*, second paragraph). Whenever we use phrases such as "low power" or "underpowered test" in this paper, we take the point of view of a scientist who assesses the power of a test to be lower than the standard they would find acceptable, given the range of effect sizes they would find interesting and plausible in the given situation.

3.3 Frequentist and Bayesian probability

The aim of the present paper is to assess the LPLC critique from the perspective of the two dominating schools of statistical inference: frequentist and Bayesian statistics. These two frameworks differ fundamentally in their understanding of probability, which is why we first need to shortly explain these differences. According to the *frequentist interpretation*, probability is the relative frequency with which repeatable processes produce outcomes (Fisher, 1935; Neyman & Pearson, 1933). This is consistent with many real-life applications of probability, such as games of chance. However, the frequentist interpretation does not allow the assignment of a probability to non-repeatable, singular events. In particular, from a frequentist perspective, we cannot assign a probability to the truth of a certain hypothesis. In contrast, the *Bayesian interpretation* treats probability as a way to quantify knowledge (e.g. De Finetti, 1974; Jaynes, 2003). Consequently, Bayesian probability *can* make statements about singular events and hypotheses.

Consider the following example: A friend tells you that she will toss a fair coin. Before the toss, she asks you about the probability of the coin showing heads. What do you answer? In this situation, the obvious answer of 50% is a valid statement in both the frequentist and Bayesian frameworks. Now, your friend tosses the coin and covers the result with her hand. What is the probability that the coin under her hand is showing heads? As a Bayesian, you still have every reason to say 50% - you know that a fair coin toss results in heads in 50% of cases, and since this particular toss happened, you gained no additional knowledge about its outcome. From the frequentist point of view, however, you find it impossible to give a meaningful answer. You cannot see through your friend's hand, but you are certain that beneath it, the coin shows either heads or tails. Thus, the only "probability" you could assign to this event would be 100% or 0%.

The same difference arises whenever the truth of scientific hypotheses is discussed. In frequentist inference, the probability of some particular hypothesis being true is not meaningful, as its truth is seen as a fixed state of the world. In contrast, Bayesian inference allows the assignment of probability to a hypothesis to quantify the existing evidence for its truth.

3.4 The original formulation of the LPLC critique

The LPLC critique originated in an attempt to transfer the logic of diagnostic medical screening to tests of epistemic inference (Colquhoun, 2014; Mayo & Morey, 2017). Its central concept is the *positive predictive value* (PPV), a term frequently used in epidemiology. There, the PPV of a test for some disease is the proportion of individuals who really have the disease among those individuals for which the test gave a positive result (Altman & Bland, 1994). The PPV is calculated, using Bayes' theorem, as

$$PPV = P\left(D|+\right) = \frac{P\left(+|D\right)P(D)}{P\left(+|D\right)P(D) + \left(1 - P\left(-|not\ D\right)\right)\left(1 - P(D)\right)}$$

where the *prevalence* $P(D)$ is the probability that any randomly selected member of the population of interest has the disease; the *sensitivity* $P(+|D)$ is the probability of a positive result, given that the patient has the disease; and the *specificity* $P\left(-|not\ D\right)$ is the probability of a negative result when the patient does not have the disease. Note that here, the use of Bayes' theorem is valid within the frequentist framework, as all involved quantities can be interpreted as relative frequencies.

The central step in the formulation of the LPLC critique is to apply the same logic to significance tests. In this framework, hypotheses take the role of „individuals“, and the truth of hypotheses takes the role of the „disease“. Sensitivity and specificity are identified with the corresponding quantities from frequentist statistics: the power of the test, typically written as $1 - \beta$, is equated with the sensitivity, while the type-I-error probability α is understood to be the complement of the test's specificity. However, the prevalence has no direct counterpart in frequentist terminology. Therefore, the quantity $P\left(H_{\text{true}}\right)$ is introduced, defined as the proportion of true hypotheses among some larger „population“ of hypotheses, for example all hypotheses tested in a certain research field. Most formulations of the LPLC critique represent $P\left(H_{\text{true}}\right)$ by the corresponding odds ratio $R = \frac{P(H_{\text{true}})}{1-P(H_{\text{true}})}$, defined as the "ratio of the number of 'true relationships' to 'no relationships' among those tested in the field" (Ioannidis, 2005, p. 696), or "the odds that a probed effect is indeed non-null among the effects being probed" (Button et al., 2013, p. 366). The PPV can then be calculated as (c.f. Ioannidis, 2005)

$$PPV = \frac{(1-\beta)P\left(H_{\text{true}}\right)}{(1-\beta)P\left(H_{\text{true}}\right) + \alpha\left(1 - P\left(H_{\text{true}}\right)\right)} = \frac{(1-\beta)R}{(1-\beta)R + \alpha}\,. \quad (\mathbf{1})$$

To complete the derivation of the LPLC critique, it is easily shown that, for any fixed value of α and $P\left(H_{\text{true}}\right)$, a decrease in power $1 - \beta$ leads to a decrease in the PPV. In the words of the LPLC critique's proponents, this relation shows that low power comes with a low "post-study probability" that the "claimed research finding is true" (Ioannidis, 2005, p. 696), and that "low power [...] reduces the likelihood that a statistically significant result reflects a true effect" (Button et al., 2013, abstract).

It should be noted that in the publications of (Ioannidis, 2005) and others (e.g. Button et al., 2013), the focus lay on the effects of low power on the evidence supplied by populations of published studies. However, the authors also invite the use of their critique to individual tests. This is seen in statements such as “the probability that a research finding is indeed true depends on [...] the statistical power of the study” (Ioannidis, 2005, p. 696) and “a study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect” (Button et al., 2013, abstract). Indeed, the LPLC critique is often also used in the assessment of single studies, and it is on this use that we focus here. With small adaptations, our arguments can also be extended to the population-level LPLC critique, though.

3.5 The LPLC critique in frequentist inference

The LPLC critique takes a peculiar position within frequentist inference. Concepts such as power and type-I-errors are central to its formulation. However, by introducing the "prevalence" $P(H_{\text{true}})$, the LPLC critique becomes decidedly "unfrequentist". Frequentist tests are designed to perform inference without any reference to the probability of hypotheses, which is even dismissed as a meaningless concept (e.g. Fisher, 1935; Neyman & Pearson, 1933). In the same vein, many recent publications describe the fallacies that arise whenever we erroneously assume that frequentist tests tell us anything about the probability of a hypothesis being true (Greenland et al., 2016; R. D. Morey et al., 2016; Wasserstein & Lazar, 2016).

Why, then, is the LPLC critique widely accepted by a predominantly frequentist field, instead of being dismissed as another such fallacy? Its similarity to diagnostic screening is very appealing, but it is easily overlooked that we cannot treat tested hypotheses like screened patients without making very strong assumptions about the nature of the scientific process. For the LPLC critique to be consistent with frequentist theory, we would need to assume that tested hypotheses are randomly sampled from some "population" of hypotheses, just as screened individuals are sampled from a population of people. This assumption is best illustrated by a variation of the classic urn model, where the urn is "filled" with true and false hypotheses (Colquhoun, 2014; Mayo & Morey, 2017). We do not know of any particular hypothesis if it is true, but we do have knowledge of $P(H_{\text{true}})$, the percentage of hypotheses in the urn that are true. We should now imagine that, when we perform scientific inference, we *randomly* draw a hypothesis from the urn, perform the significance test with a specified type-I error and power, and accept or reject the hypothesis based on the test's result. Then, the proportion of true hypotheses among all accepted hypotheses will be the PPV. However, this is clearly not a useful model of the scientific process, for a number of reasons:

- *Hypotheses are not randomly sampled:* Scientist do not randomly draw their hypotheses from some larger set of distinguishable hypotheses, all of which are equally likely to be selected. Rather, new research questions are derived from existing findings, and the results of one study change the likelihood that another study will be started.
- *It is unclear which "population" of hypotheses should be considered:* The proponents of the LPLC critique propose that $P(H_{\text{true}})$ should reflect the number of true hypotheses in the research field (Button et al., 2013; Ioannidis, 2005). But how do we define the field? E.g., as psychologists, should we consider all possible hypotheses within psychology, or only within our discipline or subdiscipline? Do we consider all hypotheses that have ever been tested in this field, or only recent hypotheses, or maybe all hypotheses that could possibly be tested in the future? Do we consider even the most absurd hypotheses, or do we restrict the population to "reasonable" ones? If the latter, how do we define "reasonable"? These questions demonstrate that the LPLC critique suffers heavily from the so-called reference class problem (Hájek, 2007; Mayo & Morey, 2017). The choice of an appropriate reference class of hypotheses must ultimately be based on the assumptions and goals of the individuals performing or assessing the analysis, introducing an element similarly subjective as the priors elicited in Bayesian analyses.

- *$P(H_{true})$ cannot be known or estimated:* The situation implied by the LPLC critique is rather peculiar: we do not know of any particular hypothesis whether it is true, yet we do know the proportion of true hypotheses in the field. It is not clear how researchers could have such substantial prior knowledge about hypotheses that have not yet been tested, without resorting to such Bayesian solutions as expert knowledge or subjective belief.

But even if we solved (or ignored) all of these issues, the PPV would still not give us the probability that any particular significant finding is true. Although the PPV has been equated with "the post-study probability that [a research finding] is true" (Ioannidis, 2005, p. 696), the frequentist probability of any particular hypothesis being true is 100% or 0%, independent of the test's result or its power. Rather, the PPV is the probability that a *randomly selected* hypothesis that produced a significant result is true. Mistaking the latter for the former is similar to the common misunderstanding that a given 95% confidence interval contains the true parameter with a probability of 95% (when it is rather the case that, in the long run, 95% of confidence intervals will contain the true parameter; R. D. Morey et al., 2016).

Ultimately, these problems arise because frequentist inference pursues a different objective than the one implied by the LPLC critique. Significance tests have been designed with the intention to achieve optimal error control in repeated decision situations, without any consideration of prior information. We run into inconsistencies, though, when we demand that significance tests inform us about the probability that claimed findings are true.

3.6 The LPLC critique in Bayesian inference

Accepting the Bayesian interpretation of probability makes the LPLC critique much simpler - we do not need to identify $P(H_{\text{true}})$ with the proportion of true hypotheses in some hard-to-define population. Instead, we may define it as the "proper" prior probability that the hypothesis in question is true, reflecting all available knowledge. We may then also interpret the PPV as the posterior probability that the hypothesis is true after observing the significant result, $PPV = P(H_{\text{true}}|\text{sig.})$. It is mathematically correct to say that lower power leads to lower values of $P(H_{\text{true}}|\text{sig.})$. However, this relationship should not lead us to dismiss significant findings from underpowered tests entirely. Rather, it prompts us to determine exactly how much (or little) evidence the results are able to provide (see also Wagenmakers et al., 2015).

Figure 1 illustrates the posterior probability $P(H_{\text{true}}|\text{sig.})$ after a test significant at $\alpha = 0.05$, for different prior probabilities $P(H_{\text{true}})$ and different levels of power. We can see that $P(H_{\text{true}}|\text{sig.})$ does decrease with decreasing power. However, whether low power makes the result not credible depends heavily on its prior probability, and on what values of $P(H_{\text{true}}|\text{sig.})$ we consider credible or not credible. If, for example, we pre-assign a prior probability of 50% to the hypothesis, significant results from tests with power levels of 0.80, 0.50, and 0.20 provide posterior probabilities of 94.1%, 90.9%, and 80.0%, respectively. Should the difference between 94.1% and 90.9% prompt us to accept the result of the test with the conventional power of 0.80, but to dismiss the result of the test with the low power of 0.50? And can we dismiss the significant result of a test with the very low power of 0.20, given that it provides the (perhaps surprisingly high) posterior probability of 80%? The answers to these questions depend on the thresholds that we deem appropriate, and on the costs of wrong decisions. It should be noted, though, that in contrast to frequentist

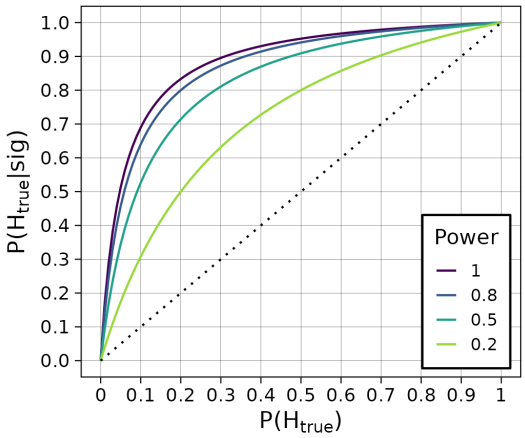

FIGURE 3.1: The relationship between prior probability $P(H_{\text{true}})$, power, and posterior probability $P(H_{\text{true}}|\text{sig.})$ after a significant result (α=0.05) is obtained.

significance testing, Bayesian inference does not demand making clear-cut (i.e. above or below a significance threshold) decisions between hypotheses.

The relative effect of power on the posterior probability changes with the prior probability. What, then, is an appropriate value of $P(H_{\text{true}})$? Despite developments regarding objective priors on the level of model parameters (Bayarri et al., 2016; Jaynes, 2003), there is no generally agreed-upon way to assign objective priors to hypotheses themselves. A common choice is 50%, implying prior indifference towards truth and falsehood of the hypothesis (Berger, 2003). But this prior might be inappropriate when the tested hypothesis appears very plausible or implausible. In such cases, it is inevitable that $P(H_{\text{true}})$ incorporates personal experience or subjective opinion, and that different stakeholders assign different priors to the same hypothesis.

Bayesian inference is often based not on the posterior probability itself, but on the *Bayes Factor* (*BF*; Kass & Raftery, 1995). The *BF* is the ratio of the probability of the observed data under the assumption that the hypothesis is true, to the data's probability under the assumption that the hypothesis is false. It plays an important role in updating the hypothesis' prior odds to its posterior odds after observing the data:

$$\frac{P(H_{\text{true}}|\text{Data})}{P(H_{\text{false}}|\text{Data})} = \underbrace{\frac{P(\text{Data}|H_{\text{true}})}{P(\text{Data}|H_{\text{false}})}}_{BF} \cdot \frac{P(H_{\text{true}})}{P(H_{\text{false}})}.$$

Importantly, the *BF* is independent of $P(H_{\text{true}})$, and can therefore be interpreted as an objective measure of how strongly the data supports the tested hypothesis. In a fully Bayesian analysis, the *BF* would be determined by assessing the probability of the full data under both H_{true} and H_{false} (Wagenmakers et al., 2015). However, if we reduce the data to the outcome of a significance test (meaning that the only information we use is whether the test was significant or not), we obtain

$$\frac{P(H_{\text{true}}|\text{sig.})}{P(H_{\text{false}}|\text{sig.})} = \underbrace{\frac{P(\text{sig.}|H_{\text{true}})}{P(\text{sig.}|H_{\text{false}})}}_{BF_{\text{sig}}} \cdot \frac{P\left(H_{\text{true}}\right)}{P\left(H_{\text{false}}\right)}$$

where BF_{sig}, the Bayes Factor obtained through a significant result, is given by

$$BF_{\text{sig}} = \frac{P(\text{sig.}|H_{\text{true}})}{P(\text{sig.}|H_{\text{false}})} = \frac{1-\beta}{\alpha}\,.$$

For a given α, BF_{sig} is a simple linear function of the power $1-\beta$. For the typical α of 0.05, the threshold of $BF_{sig} = 3$, widely seen as indicating moderate support for the hypothesis (Kass & Raftery, 1995; Keysers et al., 2020), is obtained for $1-\beta = 0.15$, a power level likely to be judged as extremely low by most scientists. A BF_{sig} of 10, the conventional threshold for strong evidence, is obtained for $1-\beta = 0.50$, a power level that is still far removed from the typically desired level of 0.80. Clearly, these observations should not lead to the conclusion that experiments do not have to be designed to achieve high power - after all, underpowered tests are unlikely to produce significant findings to begin with. However, we expect these examples to challenge scientists' intuitions about whether and to what extent low power devalues those results that have already been obtained.

It should be emphasized again that a fully Bayesian analysis would base inference on the full data (that is, the BF), not on the dichotomous outcome of a significance test (i.e., the BF_{sig}). However, there is no straightforward relationship between frequentist power and the BF. Here, we constrained our attention to BF_{sig} to facilitate the evaluation of significance tests from a Bayesian perspective. See Bayarri et al. (2016) for more details on the synthesis of frequentist and Bayesian inference.

3.7 Power and questionable research practices

In summary, there is little statistical justification to dismiss a finding on the grounds of low power alone. However, everything written above relied on the assumption that the significant result in question was obtained in an „honest way", in the sense that the reported test was the only test performed to assess the hypothesis (or, if not, that this is transparently communicated and that appropriate corrections for multiple comparisons have been performed). But significant results reported in psychological science are often produced by questionable research practices (QRPs), such as selective reporting and p-hacking (e.g. John et al., 2012). Moreover, it has been suggested that the results of studies reporting low-powered tests are more likely to be "tainted" by QRPs than those of studies with higher power (Bakker et al., 2012; Schimmack, 2012). The reasoning behind this is that researchers who do not obtain a significant result will, with a certain probability, use QRPs to nevertheless produce and report "publishable" significant findings. Researchers who conduct underpowered studies are more likely to obtain insignificant results, and thus to be tempted to use QRPs, than researchers who conduct appropriately-powered studies. Crucially, though, this reasoning holds only in cases where the alternative hypothesis is, indeed, *true*. When the alternative hypothesis is *false*, low- and high-powered studies have the same probability $(1-\alpha)$ of producing non-significant results, and are at the same risk of being affected by QRPs. Thus, maybe surprisingly, low power

only increases the likelihood that a significant result was produced by QRPs in those cases where it reflects a true finding anyway.

Does this mean that QRPs are no issue after all? To the contrary: QRPs vastly dilute the evidence we can gain from studies of low and excellent power alike. For illustration, consider the following two-step model for the generation of research "findings". First, a researcher conducts the originally planned hypothesis test. If this produces a significant result, it is reported as such. If the test is non-significant, the researcher will, with a certain probability Q, use QRPs to produce and report a significant result anyway. Now, when a researcher reports a significant result, it can either reflect a questionable or non-questionable finding, depending on whether QRPs where used or not; and it can either be a true positive (TP) or a false positive (FP), depending on whether the tested hypothesis is true or false. The resulting four possible events have the following probabilities:

$$\begin{aligned} P(\text{Non questionable TP}) &= (1-\beta)\cdot P\left(H_{\text{true}}\right) \\ P(\text{Non questionable FP}) &= \alpha \cdot P\left(H_{\text{false}}\right) \\ P(\text{Questionable TP}) &= Q\cdot\beta\cdot P(H_{\text{true}}) \\ P(\text{Questionable FP}) &= Q\cdot(1-\alpha)\cdot P(H_{\text{false}}) \end{aligned}$$

To demonstrate the effects of QRPs on evidence, it is illustrative to consider the Bayes Factor produced by a reported significant finding,

$$BF_{\text{sig.rep}} = \frac{P(\text{Non questionable TP}) + P(\text{Questionable TP})}{P(\text{Non questionable FP}) + P(\text{Questionable FP})} = \frac{1-\beta+Q\beta}{\alpha+Q(1-a)}.$$

The relationship between Q and $BF_{\text{sig.rep}}$ is depicted in Figure 3.2. For all levels of power, evidence decreases sharply with increasing Q. If we assume a high probability that non-significant findings will be "embellished" with QRPs, the evidence provided by even highly powered tests is negligibly small. If, for example, we assume Q larger than 0.30, even a test with maximal power cannot produce a $BF_{\text{sig.rep}}$ larger than 3, the conventional threshold for moderate evidence.

To conclude, the use of QRPs can completely nullify the evidence gained from results. Power plays only a side role in this. A study that shows clear signs of p-hacking or selective reporting cannot be saved by a large sample size (see also Schönbrodt & Wagenmakers, 2018, for simulation results that large sample sizes are not a protective factor against most p-hacking strategies). At the same time, low power gives little reason to question the results of an otherwise well-designed, transparently reported, and perhaps preregistered, study.

3.8 The LPLC critique in the presence of publication bias

Another phenomenon that severely affects the credibility of research output is publication bias (Franco et al., 2014). It is well established that significant findings are far more likely to be published, leading to an overestimation of effects, and even the wide-spread acceptance of false theories due to spurious evidence. This relates to the application of the LPLC critique inasmuch as it might unduly increase scientists'

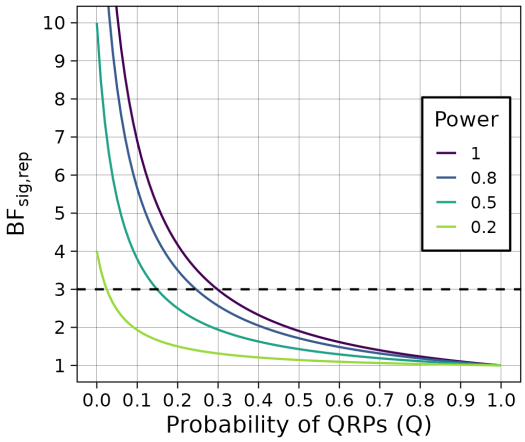

FIGURE 3.2: The effects of questionable research practices (QRPs) on the evidence provided by a reported significant result (α=0.05), as measured by the Bayes Factor ($BF_{\text{sig.rep}}$). The dashed line marks the conventional threshold $BF_{\text{sig.rep}} = 3$.

estimates of the prior probability, and subsequently also the posterior probability, of research hypotheses. These issues highlight that our belief in a hypothesis should not only be informed by the number of significant findings reported, but also on the plausibility of the hypothesis, its fit into well-established theories, and the quality of previous research.

In view of the ubiquity of publication biases, one could argue that decision-makers such as editors and reviewers should apply a "preventive" LPLC critique, denying publication to studies whose tests are, by some standard, considered underpowered. The goal of such a policy would be to prevent an "overflow effect", where false positives from a large number of small sample studies (which are quicker and cheaper to conduct than large sample studies) lead to the establishment of spurious new theories. This is an arguable position, but two crucial points should be noted: First, this would necessitate an elaborate and clearly communicated policy to decide when power will be considered too low: Given the whole range of possible effect sizes of interest, is the power of a study's tests so low that we would not want studies like this to be published, for preventive reasons? Second, this form of the LPLC critique would inevitably prioritize the strict execution of the policy over the thorough assessment of the evidence that a single study can provide. In our opinion, the issue of publication biases is better addressed by recent developments ensuring that all research findings are published, such as preregistration, registered reports, and an increased acceptance of null findings (Nosek et al., 2012, 2018).

3.9 Practical recommendations for a principled critique of low power

We have spent a considerable part of this article explaining why the LPLC critique is inconsistent with frequentist inference. However, we have also demonstrated that it is partially valid when we adopt a Bayesian perspective. Why should the applied researcher care about the inconsistencies of the critique in frequentist inference, if it is so easily "saved"? This question is especially salient now that Bayesian statistics are becoming more prevalent in psychology. However, central decision-makers such

as journal editors, reviewers, and grant committees are still binding researchers to the quality criteria prescribed by frequentist inference, such as type-I error control, correction for multiple comparisons, and rejection of optional stopping. Arguably, the increasing popularity of Bayesian statistics will ultimately oblige scientific fields to discuss and reassess whether statistical inference should be guided by frequentist or Bayesian principles, or a mixture thereof. Until then, however, the "worst of both worlds" would be if decision-makers obliged scientists to fulfill frequentist criteria, only to then judge their work on grounds that are incompatible with those same criteria.

In the following, we describe several steps that will help decision-makers such as reviewers to critique results from ostensibly underpowered studies in a principled, fair, and constructive manner. Scientists who find their work subjected to the LPLC critique should be allowed to demand that decision-makers follow these steps, and openly communicate their considerations.

(0) Acknowledge that any critique of low power depends on subjective aspects Statements about the credibility of results must, at least to some degree, depend on subjective beliefs and knowledge (e.g. estimates of prior probabilities and effect sizes). We suggest that you be open about the subjective aspects of your reasoning.

Example statement: *"I have tried to consider all available objective information in my assessment. However, the following estimates are still partially based on my subjective experience, knowledge, and reading of the literature."*

(1) Acknowledge that your critique comes from a Bayesian point of view If the reviewed work is to be solely assessed on frequentist criteria, there is no space for invoking the LPLC critique. In other cases, you can invoke it from a Bayesian point of view. Then, however, you should transparently acknowledge this.

Example: *"In my opinion, the present test(s) was(were) underpowered to detect the effect(s) of interest. I am aware that questioning significant results on the basis of low power alone is not consistent with the goals of significance tests. However, from a Bayesian point of view, I have the following concerns... ."*

(2) Explain the reasons why you think the study was underpowered, and what power you would find appropriate Any analysis can be called underpowered by assuming that the effect of interest is arbitrarily small, or by demanding that power be arbitrarily close to 1. For constructive feedback to the researchers, specify the minimal effect size of interest that you would expect in the investigated situation, the power resulting from this effect size and the study's sample size, and the power that you would deem acceptable in this situation.

Example: *"Based on the literature, I would expect that the size of the effect, if it really existed, would not be much larger than d = 0.3. The authors report a two-sample t-test with a sample size of 15 observations per group. This test would have a power of around 0.20, which I consider too low. In this situation, I would be willing to accept a power of 0.80 or larger."*

(3a) If you are concerned that the current result came about by the use of questionable research practices, communicate that concern Sometimes you might suspect that a study's significant finding results from the use of QRPs. Low power might be

one of the factors that led you to this suspicion, as the analysis had a low prior probability to find the effect it claims to detect. However, there should also be other causes for this concern, such as signs of selective reporting, abuse of researcher degrees of freedom, lack of or imprecise preregistration, etc. If this is the case, it is important that you clearly explain the reasons for your concerns. You might also name your conditions for accepting the current results.

Example: *"I am concerned that the reported result could have been obtained through questionable research practices, especially given that the presented analysis had such a low probability to detect an effect even if it were there. My concerns are further corroborated by [further evidence for QRPs]. I kindly ask the authors to provide further evidence for their claims. As it is now, a preregistered replication would be necessary to convince me of this result."*

(3b) If you are not concerned about questionable research practices, assess the result in view of all available information In the absence of concerns about QRPs, a significant result from an ostensibly underpowered test should be assessed at face-value. At this stage, you should openly discuss if, in your view, the result provides convincing evidence for the truth of the investigated hypothesis. Here, you could also make use of **formula 1** to assess the posterior probability.

Example statements: *"Given previous results, I would say that this hypothesis had a 50% chance of being true. With a power of about 0.5, the test was certainly not optimally powered. However, its significant result still gives the hypothesis a posterior probability of about 90%. This is enough evidence to make this result worthwhile for future research."*

"I would give this research hypothesis only a 10% chance of being true. The significant result from the present test, which I assume to have had a power of 0.2, raises this probability only to about 30%, which is not enough to convince me. With a power of 0.80, a significant result would raise the posterior probability to above 60%, which would make me consider it further."

"I am highly skeptical of the claims the authors are making, as they are not at all supported by previous research. I would give the authors' research hypothesis a prior probability of being true of at most 1%. The significant result from the present test, which I assume to have had a power of 0.20, raises this probability only to about 4%. But, to be fair, even the significant result from a perfectly powered test would give this hypothesis a posterior probability of only $\sim$ 17%. In my opinion, the authors need to provide much more evidence, such as additional experiments with more stringent tests, to make the claims they are making."

3.10 Conclusion

There is no question that the ubiquity of underpowered tests is a pressing problem for psychological science. However, we consider it crucial that low power is criticized for the correct reasons. Whenever a study can be constructively critiqued *before* its implementation, low power should be pointed out, and improvements demanded. However, as we have outlined in this paper, it is less straightforward to justify the critique of low power *after* a positive result has been obtained. Most importantly, low power should not be used as a proxy concern when there are deeper concerns about the trustworthiness of the reported results, and the possible use of QRPs. When suspicions about QRPs arise, they should be directly communicated as such, whenever possible. This way, it is ensured that issues of scientific integrity and transparency are not mistaken for issues of statistical power.

Acknowledgements

We would like to thank Angelika Stefan for valuable feedback on an earlier version of this manuscript.

Code availability

R code to reproduce the figures can be found on OSF: https://osf.io/7et23 .

Part II

Bayesian Cognitive Modeling

Chapter 4

Using reinforcement learning models in social neuroscience: Frameworks, pitfalls, and suggestions of best practices

Abstract

Recent years have witnessed a dramatic increase in the use of reinforcement learning (RL) models in social, cognitive and affective neuroscience. This approach, in combination with neuroimaging techniques such as functional magnetic resonance imaging, enables quantitative investigations into the latent mechanistic processes underlying social decision-making. However, increased use of relatively complex computational approaches has led to a range of misconceptions and imprecise interpretations of model outcomes. In this article, we present a comprehensive framework for the examination of (social) decision-making with the simple Rescorla-Wagner RL model. We then discuss common pitfalls in the application of this approach and provide practical suggestions. In particular, with simulation studies, we unpack the functional role of the learning rate and pinpoint what could easily go wrong when interpreting differences in the learning rate. Then we discuss the inevitable collinearity between outcome and prediction error in RL models and give suggestions of how to justify whether the observed neural activity is related to the prediction error rather than outcome valence. Finally, we suggest model validation using posterior predictive check is a crucial step after model comparison, and we articulate employing hierarchical modeling that combines group- and individual-level variance for parameter estimation. We aim to provide simple and scalable explanations and practical guidelines for employing RL models in order to assist both beginners and advanced users in better implementing and interpreting their model-based analyses.

4.1 Introduction

Computational modeling has gained increasing attention in the field of cognitive neuroscience over the past decade. It formulates human information processing in terms of basic principles of cognition, defined with latent variables in formal

This chapter is published as Zhang, L., **Lengersdorff, L. 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

Zhang L. and Lengersdorff L. L. contributed equally to this paper.

mathematical notations (Forstmann & Wagenmakers, 2015; Lewandowsky & Farrell, 2010, Table 4.1). One striking advantage of employing computational modeling is that it enables the mechanistic interrogation of trial-by-trial variations by assuming underlying cognitive processes. For classic methods based on trial summary statistics (e.g. t-test, ANOVA, linear regression), these trial-by-trial variations remain inaccessible. In addition, computational modeling requires the precise and explicit specification of parameters and variables that drive behavior. These explicit specifications allow researchers to perform stricter examinations of whether these models hold with empirical test, as opposed to more qualitative models that are much harder to be precisely confirmed. Furthermore, model-based neuroimaging approaches allow for computations on how those model-specified decision variables are implemented in the brain (J. D. Cohen et al., 2017; J. P. Gläscher & O'Doherty, 2010; O'Doherty et al., 2007). In the field of social neuroscience, the reinforcement learning framework (RL; D. Lee et al., 2012; Sutton & Barto, 2018), among other modeling frameworks (e.g. Friston & Kiebel, 2009; Ratcliff & McKoon, 2008), has been widely applied and implemented in various studies involving learning and decision-making in social contexts (e.g. Behrens et al., 2008; Burke et al., 2010; Hampton et al., 2008; Lindström, Bellander, et al., 2019; Lindström, Golkar, et al., 2019; Lockwood et al., 2016; Zhang & Gläscher, 2019) (for reviews, see Konovalov et al., 2018; Lockwood & Klein-Flügge, 2019; Olsson et al., 2020; Ruff & Fehr, 2014). Performing and interpreting computational modeling, though, comes with many challenges and potential pitfalls, especially to researchers who are new to computational approaches. For one, cognitive and social neuroscientists do not necessarily have a formal training in computational modeling, which involves multiple steps that require programming as well as quantitative skills (e.g. statistics, calculus and linear algebra; Lewandowsky & Farrell, 2010; Wilson & Collins, 2019). These skills may not always be in the core of cognitive neuroscience curricula nor the recruiting requirements (though this is currently changing at rapid pace). Furthermore, the use of RL frameworks to uncover cognitive process is still in its early stages in social neuroscience (Konovalov et al., 2018; Lockwood & Klein-Flügge, 2019; Olsson et al., 2020; Ruff & Fehr, 2014), and techniques and methods have not yet been formalized. Given these challenges, it is crucial to understand key concepts and components of RL and computational modeling before fully embracing it. In fact, we have extensively encountered with the challenges ourselves, which was one of the reasons motivating us to write this paper (see About the Authors).

In this Tools of the Trade piece, we aim to provide an easy-to-follow tutorial on the best practice of employing and interpreting RL models. We will first provide a clear and comprehensive explanation of the RL framework using the Rescorla–Wagner model (Rescorla & Wagner, 1972). Then we will pinpoint several common misconceptions and pitfalls when applying and interpreting RL models and provide suggestions of how to avoid them (Table 4.2). We will additionally discuss several practical considerations when designing RL tasks and applying RL models (Table 4.3). This tutorial is meant to be helpful to individuals who are interested in incorporating RL models into their own research, for beginners and advanced users alike. Relevant fields include, but are not limited to, social learning, social decision-making and social neuroscience. The purpose of this paper is two-fold. First, similar to other emerging fields, given the multistep nature of computational modeling, there is a large number of researcher degrees of freedom (Simmons et al., 2011; Wicherts et al., 2016) when using RL models. Here, we provide suggestions and workflows with detailed examples, with the goal to facilitate appropriate interpretation of decision

Cognitive modeling	An approximation to cognitive processes for the purposes of explanation and prediction, defined with latent variables and free parameters.
Free parameter	Model parameters to be estimated with model fitting techniques given the observed data. It does not refer to parameters that could be freely determined by the researcher (termed as fixed parameters).
Simulation	Using a known model and known parameters to generate fake/synthetic behavioral data. Simulations could be performed before data collection, and simulated data could be analyzed in the same way as real data.
Model comparison	Deciding on which model best balances goodness-of-fit and complexity, in order to examine models' generalizability and avoid overfitting. Typically, models are compared with information criteria and/or cross-validation.
Winning model	The candidate model that best balances goodness-of-fit and complexity, resulted from model comparison.
Model validation	Examining whether the winning model actually captures important features and patterns in real data. This matters because model comparison gives only relative information (whether A is better than B), whereas model validation provides absolute information (how good is A).
Hierarchical modeling	A parameter estimation technique that combines both the group-level information and the individual-level variation, by assuming an overarching group-level distribution over individuals.
Model-based fMRI	Assessing how and where model-derived computational variables (e.g., value, prediction error) are implemented in the brain using functional magnetic resonance imaging (fMRI). In practice, model-based fMRI implies the inclusion and statistical inference of a parametric regressor with model variables in the fMRI first-level design matrix.
Parametric regressor	A regressor (predictor) that tests variability in the strength (e.g., magnitude of value or prediction error) of neural responses in general linear models (GLMs). Parametric regressors commonly take on continuous values as opposed to binary values (e.g., 0 or 1).

TABLE 4.1: Glossary

variables and model-identified brain networks. Second, as social neuroscience moves forward, it is important to go beyond ad hoc repertoire of somewhat crude summary statistics toward a mechanistic understanding of neural computations with the help of computational models. This can be achieved only with a detailed and accurate comprehension of the computational algorithms behind the models and with rigorous methodological considerations. Lastly, in order to facilitate the implementation, all our analysis scripts and example dataset are openly available online under the GitHub repository (https://github.com/lei-zhang/socialRL). It is worth noting that we intend to provide neither a thorough overview of reinforcement learning nor an introduction on how to perform computational modeling and model-based analysis. For a more in-depth review of RL and computational modeling, we refer the interested readership to other excellent reviews that discuss this topic (Daw, 2011; Lockwood & Klein-Flügge, 2019; Wilson & Collins, 2019).

4.2 The Simple Reinforcement Learning Framework

The reinforcement learning (RL; Sutton & Barto, 2018) model is perhaps the most influential and widely used computational model in cognitive psychology and cognitive neuroscience (including social neuroscience) to uncover otherwise intangible latent decision variables in learning and decision-making tasks. Broadly speaking, it describes how an agent (e.g. a human participant) interacts with the uncertain external world (e.g. experimental settings) by using the feedback from the environment (e.g. monetary reward) to form internal values (e.g. participants' expected reward) of the actions or decisions it can take. Note that the value here is not an objective measure, but instead, is an internal decision variable that is assumed by the RL model. Here, the agent's action will result in feedback from the environment, which in turn will lead the agent to update the value of the action previously performed. For example, an agent's action that leads to positive feedback will be 'reinforced,' which means it will result in a higher chance of repeating that particular action. By contrast, an action that leads to negative feedback will be devalued, and subsequent similar responses are made less likely (Thorndike, 1927). A typical experiment under this RL framework unfolds as follows (Rangel et al., 2008): (i) two choice alternatives (e.g. abstract symbols) are presented, (ii) a decision is made between the alternatives, and (iii) a feedback (e.g. a monetary reward) is delivered based on the decision. Typically, the link between choice alternatives and feedback is probabilistic: one alternative comes with a high(er) chance of a reward, the other one with a complementary low(er) chance. Hence, the RL model is useful in a variety of paradigms where there is a probabilistic action–feedback association (from hereafter we will refer to this probabilistic feedback as the 'reward schedule').

Reinforcement learning encompasses an entire family of models. In its simplest form, the Rescorla–Wagner model, observations are explained by Pavlovian conditioning (Rescorla & Wagner, 1972). It has been applied to, for example, category learning (Ashby & Maddox, 2011; Poldrack & Foerde, 2008), Pavlovian and instrumental learning in reward and punishment conditions (Daw et al., 2006; Dolan & Dayan, 2013; J. Gläscher et al., 2010; Swart et al., 2017), as well as fear conditioning (Koizumi et al., 2017; Lindström et al., 2018; Norbury et al., 2018). In the field of social neuroscience, the RL model has been applied to studies that examine learning 'for' others (Lockwood et al., 2016, 2019), learning 'from' others (Behrens et al., 2008; Burke et al., 2010; Hill et al., 2017; Lindström, Bellander, et al., 2019; Lindström, Golkar, et al., 2019;

Pitfall	Suggestion
High learning rate infers fast learning, hence is more optimal than low learning rate.	The learning rate (α) quantifies the extent to which the prediction error is integrated into the value update in reinforcement learning (RL) models. High learning rate indicates fast value update that relies on only recent reward history, whereas low learning rate suggests gradual value update that carries long-lasting effect of outcomes. An "optimal" learning rate can be identified only in combination with the inverse temperature (τ). However, there is no generically optimal combination between α and τ; instead, the optimal combination is affected by the reward schedule, number of trials, the presence of reversals, and so on.
Nucleus accumbens (NAcc) encodes reward prediction error, not outcome valence.	In RL models, reward (R) and prediction errors (PE) are by definition positively correlated. However, the negative correlation between PE and value signal (V) is often overlooked, and these two theoretical subcomponents (i.e., R and V) of PE are, in fact, crucial to assess the neural substrates of PE. To qualify as a region encoding the PE signal, activities in NAcc ought to covary positively with the actual outcome (i.e., R) and negatively with the expectation (i.e., V).
Model comparison selects the winning model and validates model performance.	Model comparison is helpful in picking the best model, but it provides merely relative performance among candidate models. To validate model performance, one needs to examine whether the winning model's posterior prediction is able to reproduce key features of the observed data. This procedure is called posterior predictive check (PPC). To perform PPC, let the model generate observations (e.g., choices) from the joint posterior densities of model parameters, and then assess whether the generated data could capture the behavioral pattern (e.g., choice accuracy) as in the behavioral analysis. Unsuccessful PPC is as valuable as successful ones, because they may help falsify a model construction and eventually facilitate model development.

TABLE 4.2: Potential pitfalls and suggestions of best practices

Concern	Consideration	Example
What is a social RL task?	There are commonly two variations: either reward learning in social contexts (e.g., learn to expect monetary reward for a social partner), or social feedback learning (e.g., learn to expect social status or social evaluation).When the goal is to compare differences between social feedback and non-social feedback, we suggest matching them as closely as possible (e.g., salience, preference).	Lockwood et al. 2016; Will et al. 2017
How to initialize values in RL models for a two-option task?	As a rule of thumb, we suggest initializing values to lie between the two possible outcomes. For example, if outcomes are win (+1) and loss (-1), we suggest using 0; if outcomes are win (+1) and neutral (0), we suggest using 0.5. Note that the range of values in RL models is determined by the range of outcomes.	Wilson and Collins 2019
Is it possible to use faces as stimuli in social RL tasks instead of abstract symbols?	Faces could be used as stimuli in the same way as abstract symbols/fractals, but it is advised to be cautious to set the initial values because participants might have a priori preferences. We suggest estimating the initial value as a free parameter for each face stimulus or face category (e.g., faces of high vs. low attractiveness).	Chien et al. 2016
How many trials are required to obtain stable parameter estimates for a two-option task?	The number of trials needed to obtain stable parameter estimates depends on the reward schedule (e.g., 85:15 or 70:30). We suggest using simulation to decide the number of trials. Besides, hierarchical model estimation is preferred to obtain more stable parameter results.	Ahn et al. 2017; Valton et al. 2020; Wilson and Collins 2019
Could social RL tasks use more than two choice options?	(Social-)RL tasks are not defined by the number of choice options. Commonly, the number of options is between one and four.	Daw et al. 2006; Will et al. 2017
How to determine the range of the learning rate in RL models?	The learning rate, by definition, is between 0 and 1.	Sutton and Barto 2018
How to determine the range of the Softmax temperature in RL models?	The theoretical range of the Softmax temperature parameter is $[0, +\infty)$, yet in practice, we suggest introducing an upper limit to avoid unstable model estimation. A reasonable range is [0, 10].	Sutton and Barto 2018
Is the learning rate static across trials, or dynamically adapting along the course of the experiment?	The learning rate does not necessarily have to be constant. But in the case of the Rescorla-Wagner model (and related models), the learning rate is indeed static. A dynamic learning rate, however, is possible when other types of models are applied. Note that the interpretation of the learning rate we discussed in the main text is independent of this constant vs. dynamic property	C. Mathys et al. 2011
Does it provide additional insight to fit separate learning rates, for positive and negative feedback, respectively?	It is straightforward to extend the standard Rescorla-Wagner model with dual learning rates. However, whether the dual-learning-rate model could provide more insight than the standard Rescorla-Wagner model depends on model comparison results.	Den Ouden et al. 2013; Hauser et al. 2015
Is it possible to use RL models in the absence of choice data?	At least some sort of data is needed to perform model estimation. For example, skin conductance response (SCR) has been used to fit RL models in associative fear learning tasks, where choice data was not available.	Li et al. 2011
What are possible the ways to design a reversal learning task?	Two aspects need to be considered when designing a reversal learning task: the number of reversals (once or twice) and how often the reversals occur (after 10 trials or after 8–12 trials). A drifting reward probability could also be applied.	Den Ouden et al. 2013; J. Gläscher et al. 2009; Roy et al. 2014

TABLE 4.3: Food for thought on designing social reinforcement learning tasks and applying RL models

Suzuki et al., 2012; Zhang & Gläscher, 2019) and learning 'about' others (Hampton et al., 2008; Lockwood et al., 2018; Will et al., 2017; Yoon et al., 2018; Zhu et al., 2012).

The central idea of the Rescorla–Wagner RL model (often referred to as the simple RL model; Jones et al., 2014; Seid-Fatemi & Tobler, 2015) quantifies how error-driven learning emerges after receiving an outcome. That is, the evaluation of a choice option is updated by the difference between the actual outcome and the expected outcome. In RL models, such difference is termed as the reward prediction error, and it is formulated as follows:

$$\begin{aligned} \text{Value update: } V_t &= V_{t-1} + \alpha \cdot \text{PE}_{t-1} \\ \text{Prediction error: } \text{PE}_{t-1} &= R_{t-1} - V_{t-1} \end{aligned} \tag{4.1}$$

where for the current trial $t-1$, the reward prediction error PE_{t-1} represents the difference between the actual outcome ($R_{t-1} \in \{-1, 1\}$, for negative and positive outcomes, respectively) and the expected outcome (V_{t-1}, i.e. internal value signal). This reward prediction error is then used to update the expected outcome for the next trial (V_t), scaled by a free parameter (i.e. to be estimated with model fitting) called learning rate ($0 < \alpha < 1$; static across trials in the Rescorla–Wagner model) that calibrates the impact of the reward prediction error. Note that in the simple RL model, only the value of the chosen choice option is updated, whereas the value of the unchosen option remains intact. Note also that the reward prediction error is not necessarily specific to monetary reward; instead, it could be generalized to various other types, like food and emotion.

4.3 A Principled Interpretation of the Learning Rate

Despite the mathematical definition of the learning rate parameter (equation 4.1), its practical meaning for behavior is not always straightforward. This can hinder the understanding of the model and the interpretation of the results. In this section, we thus provide a comprehensive explanation and a principled interpretation of what role the learning rate plays in the RL model, which, to our knowledge, is sparsely covered in the literature of social neuroscience.

In simple terms, the learning rate (α) is a weight parameter that quantifies how much of the prediction error (i.e. the difference between the actual and the expected outcome) is incorporated into the value update (i.e. the evaluation of the expected outcome of choice alternatives)—the higher this parameter, the stronger the weighting of the prediction error for the value update. For instance, when α is 0, the value of the chosen option is not updated at all, whereas when α is 1, the value of the chosen option is updated using the entire reward prediction error. Similarly, when α is 0.5, half of the reward prediction error is used for the value update. Given this property, the learning rate is often considered as the step size of learning (Sutton & Barto, 2018) or the speed of learning (J. Gläscher et al., 2009; D. Lee et al., 2012; Lockwood et al., 2016). In this view, the higher the learning rate, the faster the learning. Unfortunately, this is only half the story, and the interpretation can be misleading without unpacking the role of α in the RL model. We will demonstrate this below.

The learning rate indeed reflects the speed of learning. In a learning environment where the reward schedule is 75:25 (i.e. 75% probability of receiving positive outcome

and 25% probability of receiving negative feedback), a high learning rate (e.g. $\alpha = 0.9$) leads to quicker value updating, and the updated value will approximate its maximum after only two trials, if positive outcomes (e.g. monetary reward) are observed (Figure 4.1A, trials 2–3). However, this high learning rate also causes a dramatic value decrease after receiving only one negative feedback (Figure 1A, trial 7). This means that a high learning rate is helpful for obtaining faster value updates after receiving positive feedback; however, at the same time, it will also cause oversensitivity to negative feedback (in cases where the better choice alternative leads to negative feedback, i.e. rare probabilistic negative feedback). In contrast to higher learning rates, a lower learning rate (e.g. $\alpha = 0.3$) will result in slower value updating after positive feedback, but it will result in less sensitivity to negative feedback (Figure 4.1A). How does it come to such asymmetrical effects of the learning rate? This is due to the second important property of the learning rate parameter that governs how much recent outcomes carry over to the value update on the current trial. Crucially, equation 1 could be redefined as a function of the initial value and the outcome per trial:

$$\begin{aligned} V_t &= (1-\alpha)V_{t-1} + \alpha R_{t-1} \\ &= (1-\alpha)(V_{t-2} + \alpha(R_{t-2} - V_{t-2})) + \alpha R_{t-1} \\ &= (1-\alpha)^{t-1}V_1 + \sum_{i=1}^{t-1}(1-\alpha)^{t-i-1}\alpha R_i \end{aligned} \tag{4.2}$$

where V_1 is the initial value, t indexes the current trial, and the term $\sum_{i=1}^{t-1}(1-\alpha)^{t-i-1}\alpha R_i$ depicts the carry over outcome weight, formulated as a cumulative sum of each trial's outcome (R_i). Simply put, this formula describes how much outcomes in the past contribute to the current value computation. In the case of a high learning rate (e.g. $\alpha = 0.9$), only the two most recent outcomes ($t-1$ and $t-2$) contribute to the value update, and the weight on previous outcomes is strongly reduced (Figure 4.1B). This is in contrast to a low learning rate (e.g. $\alpha = 0.3$): although the impact of the current outcome is weaker relative to a high learning rate, it shows more long-lasting effects of outcomes that are received from early on ($t-1$ to $t-8$; Figure 4.1B). In short, these mathematical properties suggest that when the learning rate is high, only the most recent outcomes matter for the value update, whereas when the learning rate is low, both recent outcomes and outcomes from further back contribute to the value update. This example illustrates why a value update with a high learning rate is much more influenced by fewer trials, whereas value update with a low learning rate is not overly sensitive to negative feedback.

4.4 Is there an optimal learning rate?

Overall, a high learning rate suggests faster learning, which will be influenced most strongly by the most recent outcomes; a low learning rate indicates slower but steadier learning, which is influenced by a larger number of past trials (compared to a high learning rate). These results bring us to the question: which one is the 'better' or more 'optimal' learning rate? Which learning rate will result in more correct choices overall? Is a higher learning rate better than a lower learning rate or vice versa? This is a central question of many social neuroscience studies using computational modeling to identify differences in behavior by analyzing differences in parameter estimates.

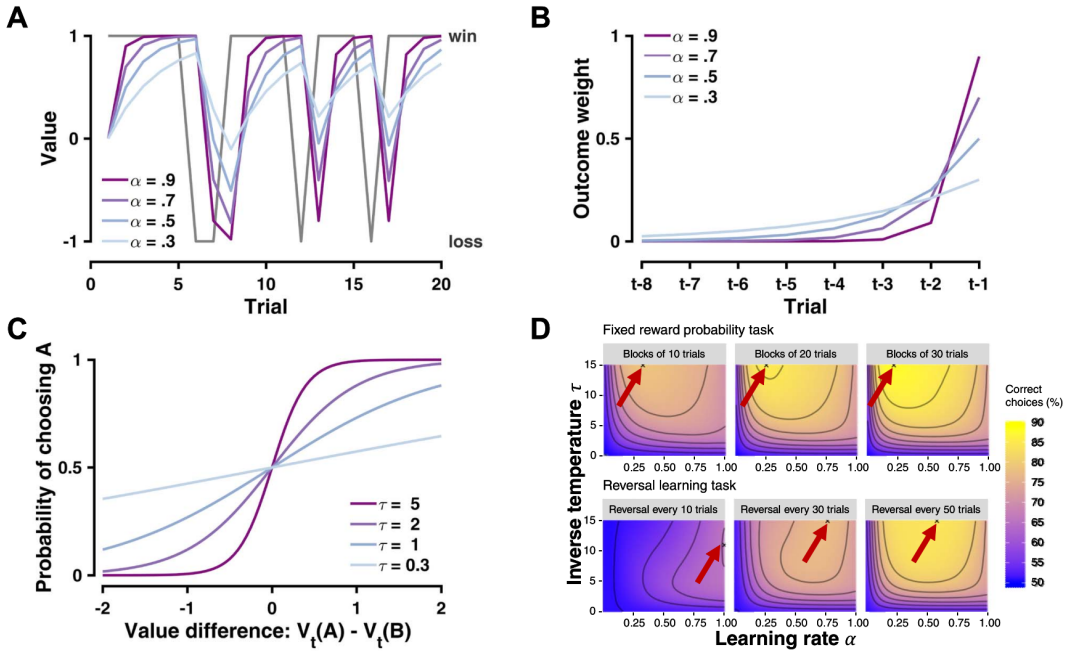

FIGURE 4.1: Comprehending model parameters of the Rescorla–Wagner model. **(A)** Effect of different learning rates on value update. **(B)** Effect of different learning rates on the weights of past outcomes. **(C)** Effect of Softmax inverse temperature on converting action value to action probability. **(D)** Effect of different combinations between learning rate and Softmax inverse temperature on choice accuracy per trial length (blocks of 10, 20 and 30 trials), for fixed reward schedule tasks (top) and reversal learning tasks (bottom). Arrows depict the 'optimal' combination that predicts highest choice accuracy for each task setup

Researchers using RL models have identified differences in the parameter estimates characterizing behaviors of healthy and psychiatric populations (e.g. Fineberg et al., 2018; Lin et al., 2012) and of individuals subjected to a pharmacological treatment vs. control individuals (Crockett et al., 2008, 2010; Eisenegger et al., 2013). In social neuroscience, for example, there is evidence that different learning parameters may characterize learning in social vs non-social contexts and in self-oriented vs other-oriented learning (e.g. Lockwood et al., 2016). In the latter case, what does this imply in terms of models testing for differences between self- and other-oriented optimal decisions and their possible implications for prosociality and its neural bases?

Here, we demonstrate that there is no generically optimal learning rate and that the 'good' or 'optimal' heavily depends on the research question and the experimental design (Behrens et al., 2007; Daw, 2013; Frank et al., 2007; Soltani & Izquierdo, 2019). More importantly, this 'optimal' behavior cannot be correctly understood without the other free parameter in the RL model, the inverse Softmax temperature. In RL models, after values are updated with the reward prediction error (equation 4.1), the next step is to utilize those values on the next trial and make a new decision. This value–choice connection is depicted by the Softmax choice rule (Sutton and Barto, 2018), which serves as the likelihood linking function that bridges the model with the observed data. In the case of choosing between two options A and B, the Softmax choice rule converts action values (e.g. $V(A), V(B)$) to action probabilities (e.g. $p(A), p(B)$): the higher the value of A, the more likely A will be selected:

$$p_t(A) = \frac{\exp\left(\tau\, V_t(A)\right)}{\exp\left(\tau\, V_t(A)\right) + \exp\left(\tau\, V_t(B)\right)} = \frac{1}{1 + \exp\left(-\tau(V_t(A) - V_t(B))\right)} \tag{4.3}$$

where p is the probability of choosing option A, determined by the Softmax function with the inverse temperature parameter ($\tau > 0$). In the case of two choice options, the Softmax function is simplified as a logistic curve, where the input is the value difference between V(A) and V(B), and the output is the probability of choosing A (equation 4.3). The inverse temperature parameter (τ) is the slope of the sigmoid curve, which measures choice consistency. When τ is small (e.g. $\tau = 0.2$), the curve is shallow, which represents more random choices (i.e. less consistent). In contrast, when τ is large (e.g. $\tau = 5$), the curve is steep, and this represents more consistent choices, in favor of the high(er) reward option (Figure 4.1C). Together, these two parameters, the learning rate (α) and the inverse temperature (τ), form the basis of the simple RL modeling in practice. In order to define 'optimal' behavior (i.e. deciding on the more rewarding option), we recommend researchers to interpret the joint parameter space of α and τ, rather than either of them alone. Whether a certain value of α is 'better' or 'worse' than another depends crucially on the value of τ and vice versa. Moreover, the optimal combination of these two parameters depends also on features of the task design. We performed a simulation study to demonstrate this point (i.e. simulating synthetic data using known parameters; see also Appendix 4.A). Specifically, when the reward schedule is fixed (e.g. 75:25 throughout the entire experiment), a relatively low learning rate ($\alpha \approx 0.25$), paired with a high inverse temperature, is optimal in choosing the more rewarding choice option (Figure 4.1D, upper panel). Higher learning rates in such stable learning environment would cause suboptimal behavior, because of the oversensitivity to negative feedback demonstrated above. However, the optimal parameter combination is different in the context of reversal learning tasks (i.e. the reward schedule reverses: the more rewarding option becomes

less rewarding and vice versa) than stable learning environment. When moderately frequent reversal exists in the learning environment (e.g. the reward schedule reverses every 30 or 50 trials), a relatively high learning rate ($\alpha > 0.6$), together with a high inverse temperature, is optimal. When reversals take place more rapidly (e.g. every 10 trials), a very high learning rate ($\alpha \approx 1$), in combination with a moderately high inverse temperature, is the optimal parameter (Figure 4.1D, lower panel).

Importantly, the simulations also illustrate that the relationship between α and the percentage of correct choices depends on the value of τ. Taking a task with fixed reward schedule (Figure 4.1D, upper panel) as example, we observe that for low values of τ ($0 < \tau < 2$), increases in α generally lead to increases in correct choices (but note that increases already become negligible at $\alpha \approx 0.25$). For higher values of τ, however, the relationship becomes non-monotonic: for lower ranges of α ($0 < \alpha < 0.25$), the relationship between α and correct choices is positive. After that, however, increases in α lead to a decline in the number of correct choices. As an example of how ignoring this complex behavior might lead to misinterpretations, imagine that a research team observes that a certain drug increases the average learning rate in such a RL task from 0.5 to 0.75. They may conclude that the drug manipulation improved participants' learning abilities. However, such a qualitative claim might be invalid without also taking the values of τ (as well as the task design) into consideration. If participants had low values for τ (< 2), then such a claim might be warranted. By contrast, if participants had very high values for τ, a principled interpretation of this result should rather lead to the conclusion that the drug manipulation gave rise to less adaptive behavior in this task. Of course, the situation would be more complex still if there was greater interindividual variation within τ or if the drug manipulation also influenced the average value of τ. In either case, a joint interpretation of parameters, for example, helped by simulations, would be especially crucial.

In summary, these results demonstrate that the practical meaning and interpretation of parameters vary between task designs, and researchers must be cautious when inferring from differences in parameter estimates to differences in actual behavior. Importantly, these results concur with our principled interpretations of the learning rate demonstrated in the last section. On one hand, when the reward schedule is stable, it is crucial for an agent to ignore rare and misleading negative feedback; otherwise, oversensitivity to negative feedback would lead to suboptimal behavior. On the other hand, when the reward schedule is volatile, negative feedback is in fact informative where the most recent outcome matters more than previous trials; hence, it is important for an agent to detect the reversal and recompute the action values. Lastly, the recommendations delivered here demonstrate the usefulness of simulation studies, which is a powerful tool that helps to deepen our understanding of the underlying computational models and to avoid misconceptions when interpreting the results. It is noteworthy that simulations could be performed before actual data collection (see Palminteri, Lefebvre, et al., 2017; Wilson & Collins, 2019, for a discussion).

4.5 Scrutinizing the neural correlates of the prediction error in the brain

Once we have obtained the individual-level parameters for our model estimations, we are able to derive trial-by-trial decision variables that can then help us probe the moment-by-moment cognitive processes when participants are engaged in RL

experiments. In the case of the simple RL model, researchers are often interested in the prediction error, alongside the subjective value of the chosen option (i.e. chosen value), which is derived from the model using individual learning rate (α) and inverse temperature (τ). When combined with functional magnetic resonance imaging (fMRI), model-based fMRI (J. D. Cohen et al., 2017; J. P. Gläscher & O'Doherty, 2010; O'Doherty et al., 2007) allows us to identify where and how the prediction error computations are carried out in the brain. In this regard, there is substantial evidence that activity in the nucleus accumbens (NAcc) exhibits a parametric relationship to the reward prediction error (O'Doherty et al. 2004; O'Doherty et al. 2003, 2017; for a review of other brain regions that are associated with PE, see O'Doherty et al. 2017), and this finding has been replicated across several studies (Behrens et al., 2008; Chien et al., 2016; Jocham et al., 2014; Klein et al., 2017; Pagnoni et al., 2002; Pessiglione et al., 2006; Zhang & Gläscher, 2019). However, it is often ignored that including purely the actual outcome (i.e. win or loss, coded as 1 or -1) in the fMRI design matrix, instead of the fine-grained trial-by-trial prediction error derived from the RL model, contributes to similar neural response in NAcc (e.g. Cools et al., 2006; Guitart-Masip et al., 2011; Klucharev et al., 2009). The question here is does NAcc parametrically encode trial-by-trial prediction error or respond merely to the outcome valence? In other words, if both prediction error and outcome valence were accompanied by activity in NAcc, how could we conclude whether NAcc is encoding prediction error, rather than outcome valence?

Outcome (R) and prediction error (PE) are often highly correlated; it should, therefore, not be surprising that both R and PE are associated with activity in NAcc. The key to delineating this collinearity is to inspect the mathematical definition of the PE, in order to justify a neural representation of its signal. PE is computed by the difference between its two subcomponents, the difference between R and V (expected outcome; equation 4.1 and Figure 4.2A). Thus, PE ought to positively correlate with R and negatively correlate with V (Figure 4.2B). The intuition here is that the larger the reward (R), the larger the difference between R and the expectation (V), hence larger PE, whereas the higher the expectation (V), the lower the positive (or negative) PE when being rewarded (or unrewarded) (see Appendix 4.B for more details). Given that the goal of model-based fMRI is to identify the neural indicators of latent computational variables, activity in the NAcc should show similar correlation patterns with R and V, respectively. In other words, to qualify as a neural basis of the PE, activity in the NAcc should covary positively with R and negatively with V (Figure 2A). For example, in one previous study (Zhang & Gläscher, 2019), we found that activities in the NAcc showed a positive relationship with R (mean effect size 3–7 s after outcome onset: 0.230, p<0.0001, permutation test) and a negative effect of V (mean effect size 3–7 s after outcome onset: -0.033, P=0.021, permutation test; Figure 4.2C). These findings also replicate previous studies that demonstrated similar patterns (e.g. Behrens et al., 2008; Jocham et al., 2014; Klein et al., 2017; Niv et al., 2012).

In summary, when assessing neural correlates of any error-like signal, we recommend a two-step procedure: first identify and then justify (see Appendix 4.C for practical details). First, identify the neural correlations of PE with model-based fMRI analysis (i.e. including PE as the sole parametric regressor). Second, justify whether the resulting brain areas are indeed associated with PE, rather than outcome valence, by considering PE's two theoretical subcomponents (i.e. actual outcome$R$ and expected outcome V; Wilson & Niv, 2015). Only when activities from the resulting brain area(s) positively covary with the actual outcome (R) and negatively covary with the

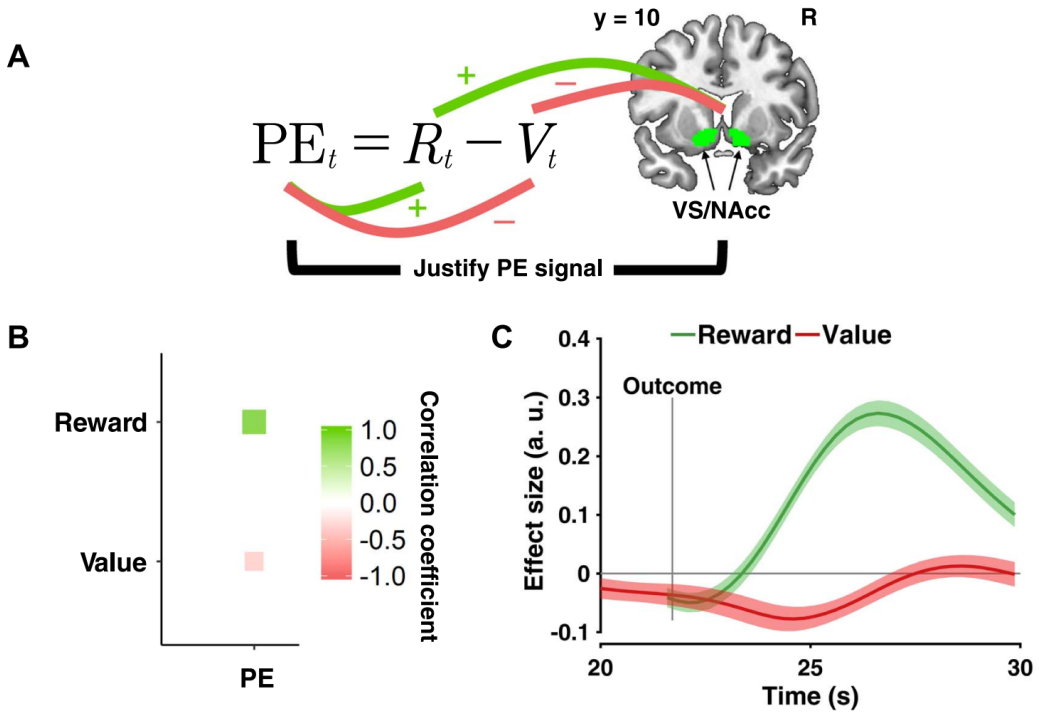

FIGURE 4.2: Justifying neural correlates of the prediction error signal. **(A)** Conceptual illustration of justifying the neural representations of the prediction error (*PE*) signal using *PE*'s two theoretical subcomponents, observed in the nucleus accumbens (NAcc). **(B)** Correlations between *PE* and its two subcomponents derived from the model. **(C)** Relationships of NAcc time series with *PE*'s two subcomponents, after outcomes are delivered. Figure adapted from Zhang and Gläscher 2019.

expected outcome (V) should that area(s) be justified as the neural basis of prediction error signaling (see also Appendix 4.D for a discussion and concern of entering $R$and PE together into the design matrix). If those activities of the brain area only have a positive relationship with the actual outcome but no correlation with the expected outcome, this area only responds to the outcome valence.

In the field of social neuroscience, it is of interest to examine a social prediction error and its neural correlates (SPE; e.g. Lockwood et al. 2016; Sun and Yu 2014; Zhang and Gläscher 2019; for a review, see Lockwood and Wittmann 2018), but sometimes the definition of the social prediction error is unclear. Is the SPE actually a reward prediction error embedded in a social context (e.g. other-oriented reward learning; Ihssen et al., 2016; Lockwood et al., 2016)? Or does the SPE indeed reflect the difference between actual and expected social feedback (e.g. suggestions or appraisal from others; Behrens et al., 2008; Will et al., 2017; Zhang & Gläscher, 2019)? In both cases the SPE is valid and has given rise to important findings. As for the neural correlates of the SPE, when the SPE is the reward PE in a social context, the justification of its neural correlates is the same as that for the reward PE. When the SPE is reflecting the predictions driven by social feedback, we recommend identifying first what the actual social feedback and expected social feedback is and then using the aforementioned criteria to test the neural correlates of the social prediction error signal, so as to prevent unnecessary inflation of ill-specified analysis and misinterpretation of the social prediction error.

4.6 Model validation is as important as model comparison

When performing computational modeling, it is rarely true that only one single model is considered to test the potential cognitive processes. More commonly, researchers (i) fit several candidate models that vary in terms of model assumption and complexity and then (ii) pit models against one another to decide on the 'winning model' (Forstmann & Wagenmakers, 2015; Lewandowsky & Farrell, 2010; Wilson & Collins, 2019). The first procedure is called model estimation and can be achieved using several model fitting techniques, such as least squares, maximum likelihood estimation, maximum a posteriori estimation and Bayesian estimation (McElreath, 2018). The second procedure is called model comparison, and it balances a model's goodness-of-fit and its generalizability (Forstmann & Wagenmakers, 2015; Gelman et al., 2013; Lewandowsky & Farrell, 2010; Wilson & Collins, 2019). This is often done with cross-validation or information criteria (e.g. the Akaike information criterion, AIC; Gelman et al. 2013; Sakamoto et al. 1986; and the widely applicable information criterion, WAIC; Gelman et al. 2013; Vehtari et al. 2016). Model-based analyses are then carried out using decision variables derived from the winning model. We argue, however, that in-between deciding on the winning model and any subsequent model-based analysis, it is necessary to validate the model as well. This is not trivial because model comparison merely considers the merits of each model's performance relative to the rest (i.e. relative scale). Thus, there is no guarantee that the winning model survived from model comparison indeed explains or predicts the behavioral effect of interest. As an example, among models whose predictive accuracy are 45, 50 and 55%, model comparison would identify the model with 55% accuracy as the best among these three candidates; nevertheless, it is still a relatively poorly performing model that predicts not much higher than chance. Therefore, it is necessary to perform model validation—examining how well the model predicts the data. Here we demonstrate the validation of a simple RL model with a posterior

predictive check (PPC), the most widely used model validation approach (Gelman et al., 2013; Levy et al., 2009; Lynch & Western, 2004).

In essence, the PPC utilizes the parameters' joint posterior distribution obtained from model estimation to generate new predictions and compares whether those predictions are able to account for effects in the observed data. The predictive distribution is defined as follows:

$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}|\theta)p(\theta|y)d\theta, \tag{4.4}$$

where

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta). \tag{4.5}$$

Here, $p(\theta|y)$ is the joint posterior distribution of model parameters given observed data, and it is obtained by model estimation techniques with the help of the Bayes' rule (equation 4.5). The predictive density, $p(y_{\text{rep}}|y)$, depicts the degree to which model-reproduced data (y_{rep}) corresponds to the actual data (y). This means a PPC assesses how the predictions of a model deviate from observed data. Importantly, any model's inability to account for the key features in behavior falsifies the corresponding model. In turn, this opens up means to improve the model (Korn & Bach, 2018; Palminteri, Wyart, & Koechlin, 2017). To perform a PPC in the context of a simple RL (Aylward et al., 2019; Frank et al., 2015; Haines et al., 2018; Steingroever et al., 2013, 2014, see also Appendix 4.E for detailed steps), we let a model generate synthetic choice data per trial and per participant with the individual-level parameters acquired from model estimation, for multiple times (e.g. 1000). Then we conduct the same behavioral analysis as we previously did with the actual data; in this case, this is how often participant chooses the more rewarding option (i.e. percent correct choices). Finally, we compare the deviance (e.g. mean squared deviation, MSD) between results from synthetic data and actual data. Typically, a PPC is performed at three levels. At the trial level (averaging across participants), a PPC examines the trial-by-trial dynamic of the choice behavior (Figure 4.3A). At the participant level (averaging across trials), a PPC assesses the individual variation (Figure 4.3B). At the overall level (averaging across both trials and participants), a PPC provides the average performance of the model (Figure 4.3C). To draw inferences, at either the trial level or the participant level, a simple correlation between model and data can be calculated to examine their association. At the overall level, a Bayesian p value (Gelman et al., 2013) could be computed to assess how much area under the posterior curve is below the actual data. When the model systematically under-/overestimates the actual data, we suggest including additional computation components (either additional steps or additional parameters) that may overcome this bias.

In the field of social neuroscience, only a handful of studies have so far employed PPC (e.g. Lindström, Bellander, et al., 2019; Lindström, Golkar, et al., 2019; Zhang & Gläscher, 2019). However, now that most cognitive models are constructed using mathematical operations, which entail the feature to generate new data, we strongly articulate that performing PPCs is indispensable when conducting computational modeling. This is especially true when Bayesian estimation is employed for fitting models (see the next section). Lacking such direct assessment between generated data and observed data might lead to less sound support of the associated cognitive

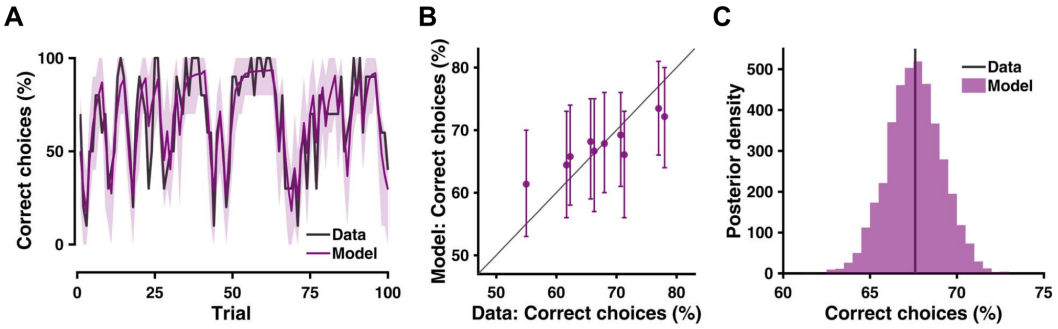

FIGURE 4.3: Model validation with posterior predictive check (PPC). **(A)** Trial-by-trial model predictions plotted against actual data. Shaded area depicts the 95% highest density interval (HDI) of the posterior distribution. **(B)** Individual's model prediction compared with actual data, in relation to the identity line. Error bars depict the 95% HDI of the posterior distribution. **(C)** Grand average model prediction across trials and participants.

processes, hence resulting in weaker implications when carrying out subsequent analyses.

4.7 Moving toward hierarchical model estimation

There are multiple ways to estimate the learning rate and the inverse temperature parameters in RL models, with some being more appropriate than others. Here, we describe why we advocate a hierarchical approach and provide several examples of toolboxes one can use to perform hierarchical modeling.

Typically, in a study, we are interested in the behavior of several subjects and possibly several groups of subjects (e.g. comparing clinical vs non-clinical populations or treatment vs control groups). The most straightforward approach is to assume that the parameters are fixed across subjects and estimate one set of parameters for the entire population. This, however, inevitably ignores the between-subject variability and could potentially lead to overstating the differences in parameter means between groups; hence, results from this approach are not generalizable to the population (Gelman et al., 2013; Maxwell et al., 2017; McElreath, 2018). A prevailing approach has therefore been to estimate a set of parameters for each subject separately (i.e. treating the parameters as random effects) and then use simple linear models (e.g. t-tests) to compare groups. The crucial problem of this method is that it tends to lead to extreme parameter values (e.g. unlikely low or high learning rates) and it inflates the population variance (Daw, 2011; Lebreton et al., 2019); thus, it is not well-suited for comparing group-level statistics.

It is therefore more accurate and stable to estimate both the individual and group level parameters simultaneously using hierarchical models (often interchangeable with multilevel model, mixed model, or partial pooling; Gelman et al., 2013; Maxwell et al., 2017; McElreath, 2018). The main benefit of this approach is that we pool data across individuals to explicitly estimate the mean and variance of the population. The group estimates then conversely constrain the parameter estimation of each individual in the group, thus avoiding unlikely extreme values. Although the advantages of hierarchical models have long been recognized (Ahn et al., 2017; Bartlema et al., 2014; M. D. Lee, 2011; M. D. Lee & Wagenmakers, 2014), it has until recently been

computationally too demanding and complicated to perform. Advances in computing power and approximation methods have led to developments of readily available tools and packages that make these previously less accessible methods easy to use. These include the hBayesDM toolbox (Ahn et al., 2017), the VBA toolbox (Daunizeau et al., 2014), the HGF toolbox (C. Mathys et al., 2011; C. D. Mathys et al., 2014) in the TAPAS software collection (https://git.io/fjUn8), the HDDM toolbox (Wiecki et al., 2013), the MFIT toolbox (https://git.io/Je0jw) and the CBM toolbox (Piray, Dezfouli, et al., 2019). All these toolboxes have optimized model specifications and considered proper parameter distributions. In fact, many of the toolboxes have already been used in social neuroscience (e.g. Fineberg et al. 2018; Hu et al. 2018; Zhang and Gläscher 2019; see Appendix 4.F for statistical considerations for using hierarchical models to compare group differences; see also Valton et al. 2020 for a review), and following the workflow provided by these toolboxes is less prone to misspecification and misinterpretation (in the case of Type II error) when applying computational modeling.

4.8 Concluding summary

In the present work, we raise the awareness of common pitfalls of employing reinforcement learning (RL) in social neuroscience and give suggestions of best practices. This is to provide a set of guidelines to accurately interpret results and to improve scientific practice when using computational modeling. With the simple Rescorla–Wagner RL model as the example, we provide a principled and detailed interpretation of the learning rate using simulation, followed by the illustration of inspecting the theoretical subcomponents of the prediction error to justify its neural correlates. Finally, we address the necessity of validating models alongside model comparison. Besides, we illustrate the consideration of performing hierarchical modeling and direct readers to available toolboxes. Note that our suggestions (except for justifying the PE signal) are neither specific to reinforcement learning nor to social neuroscience. We recommend running simulations to gain a better comprehension of parameters for all kinds of computational models, and importantly, this can be accomplished even before data collection. Note that although RL models have been shown to be a powerful tool in social neuroscience, other forms of models are insightful as well. For example, the inequality aversion model has been employed to assess separate types of inequality aversion in economic games (Gao et al., 2018; Hu et al., 2018; van Baar et al., 2019); the Bayesian belief update model has been used to study moral decisions (Siegel et al., 2018, 2019); the interactive partially observable Markov decision process (i-POMDP) has been applied to examine higher-level theory of mind during strategic social interaction (Doshi et al. 2009; Hula et al. 2018, see Rusch et al. 2020 for a review). Even under the RL framework, other types of RL models, for example, the two-step RL model (Lockwood et al., 2019) and RL model with dynamic learning rate (Piray, Ly, et al., 2019), are also applied to a range of social learning paradigms. Discussing all above models, however, is beyond the scope of this paper; instead, listing them serves as a roadmap so that interested readers will know where to start when coming across these models in their own research.

To conclude, when handled properly, reinforcement learning models can uncover insightful cognitive processes that are otherwise intangible with classic approaches in social neuroscience. It is important to minimize misconception and misinterpretation when applying reinforcement learning models in social neuroscience. These

suggestions are meant to aid future studies in interpreting and unpacking the neuro-computational mechanisms of social behaviors.

Data and Software Availability

Data and custom code to perform simulation and analysis can be accessed at the GitHub repository: https://github.com/lei-zhang/socialRL.

About the Authors

L.Z. holds a PhD in cognitive neuroscience and is the co-developer of the hBayesDM package; L.L. is a PhD candidate, holds a master's degree in psychology and has advanced training in statistics; N.M. is a PhD candidate and holds dual master's degrees in cognitive science and in mathematics; J.G. is a principal investigator in computational neuroscience; and C.L. is a professor in social neuroscience.

Acknowledgements

We thank Daisy Crawley, Yang Hu, Christoph Korn, Lukas Neugebauer, Jonas Nitschke, Saurabh Steixner-Kumar, Tessa Rusch and Antonius Wiehler for their helpful discussions and comments on the conceptualization and earlier versions of this manuscript, as well as two anonymous reviewers who greatly improved the manuscript.

Funding

This work was in part supported by the International Research Training Groups 'CINACS' (DFG GRK 1247), the Research Promotion Fund (FFM) for young scientists of the University Medical Center Hamburg-Eppendorf, National Natural Science Foundation of China (NSFC 71801110), the Ministry of Education in China Project of Humanities and Social Sciences (MOE 18YJC630268) and China Postdoctoral Science Foundation (No. 2018M633270) to L.Z.; the Bernstein Award for Computational Neuroscience (BMBF 01GQ1006), the Collaborative Research Center 'Cross-modal learning' (DFG TRR 169) and the Collaborative Research in Computational Neuroscience (CRCNS) grant (BMBF 01GQ1603) to J.G.; the Vienna Science and Technology Fund (WWTF VRG13-007) and Austrian Science Fund (FWF P 32686) to C.L.

Conflict of Interest

The authors declare no competing financial interests.

4.9 Appendix

4.A Using simulations to guide interpretation of parameters

One important aspect of learning tasks is the so-called optimal behavior. Researchers are often interested in evaluating behavioral patterns in these tasks as more or less adaptive, as "better" or "worse" for the agent. Here, "adaptive behavior" can be easily defined as choosing the option that results in more rewards and/or in fewer punishments, in other words, the choice accuracy. However, when such considerations guide the interpretation of results, researchers must take extra caution when interpreting differences in parameter estimates. Whether a certain difference in parameters indicates "better" or "worse" behavior can be heavily dependent on the task design and the values of the other relevant model parameters. Moreover, intuition can be misleading: the term "learning rate" of the parameter α of the basic RL model might imply that high values correspond to adaptive behavior. This is, however, not necessarily the case. The good message is that simulations can help us understand what different parameter configurations mean with regards to optimal behavior. To quantitatively demonstrate this, we ran a simulation study using R 3.6.1 (R Core Team, 2021). We simulated participants with known combinations of model parameters. The parameter combinations were taken from a grid spanned by the learning rate (α; minimum value/maximum value/steps = 0/1/60) and inverse temperature τ (0/15/60). Each of these virtual participants then completed a large number of trials (> 50,000) from (1) a fixed probability learning task, and (2) a reversal learning task, both with a reward schedule of 75:25. For each virtual participant, we then calculated the percentage of correct choices as the percentage of choices for the option that was associated with a higher probability for a reward. To reduce random noise due to the finite number of samples, we smoothed the resulting images with a gaussian filter (SD = 2). To illustrate the influence of the task design on the parameter-behavior-relationship, we repeated this analysis with varying task parameters. For the fixed probability learning task, we ran the simulations with block lengths of 10, 20, or 30 trials. For the reversal learning task, the block length was fixed to 200, and we varied the number of trials after which a reversal of the reward probabilities happened (i.e., reversal after every 10, 30, or 50 trials)

For the fixed probability task, we observed that:

1. The overall appearance of the plots (Figure 4.1D) changes with the number of trials within a block. Higher percentages of correct choices can be attained with longer block lengths. This appears reasonable, as participants have more opportunities to learn about the options when there are more trials

2. Both α and τ must be above a certain minimal value for participants to make correct decisions above chance level. However, once that minimal value is exceeded, the relationship between the parameters and the percentage of correct choices is not necessarily monotonic. With regards to τ, higher values result nearly always in a higher number of correct decisions. With regards to α, however, the relationship is more complex: in the lower range of values, there is a sharp increase in correct decisions with increasing values in α, until a maximum is reached. With higher values of τ, and a higher number of trials, this increase is sharper, and the maximum is reached at lower values of α. After the maximizing value is reached, further increases in α result in a moderate decline in correct decisions (e.g., Figure 4.A.1, for blocks of 20 trials)

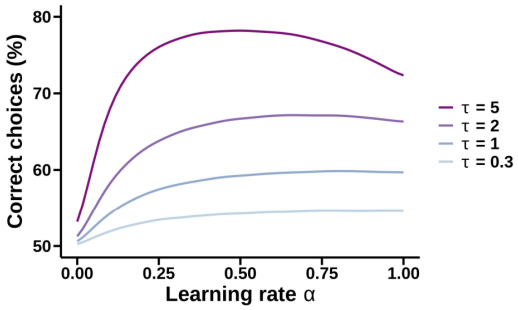

FIGURE 4.A.1: Influence of learning rate (α) on choice accuracy when paired with four levels of inverse temperature (τ). Fixed probability task for a block of 20 trials.

3. The global maximum (marked in the Fig.1D with an arrow) is attained at the highest possible value of τ (15 in these simulations), and relatively low values of α. Additional analyses allowing for higher values of τ (easily replicable with the code available online) suggest that the global maximum cannot be obtained with finite values of τ. The maximum is always attained with the highest possible value for τ, while the value for α that maximizes the output moves closer to zero as the upper bound for τ is increased.

For the reversal learning task, we observed that:

1. In general, a higher percentage of correct choices is possible in task designs where reversals happen less often (after a larger number of trials), suggesting that optimal decisions are easier in less volatile environments.

2. The relationship between parameter values and the percentage of correct decisions changes significantly with the number of trials before a reversal happens. In the task with the most volatile environment (reversal every 10 trials), the number of correct decisions increases steadily with the learning rate, and (independent of τ) the maximum is attained at the maximal value of 1. In contrast, values of τ need to be intermediate for maximal percentages of correct choices, indicating that, in very volatile environments, a certain degree of exploration is advantageous.

3. Tasks with less volatile environments (reversal after 30 or 50 trials) are more similar to those of the fixed reward probability tasks: In general, the number of correct decisions increases with increasing values of τ, and is maximized for intermediate values of α. However, it can be observed that in the reversal learning task, the optimal values of α are in general higher than in the fixed reward probability task.

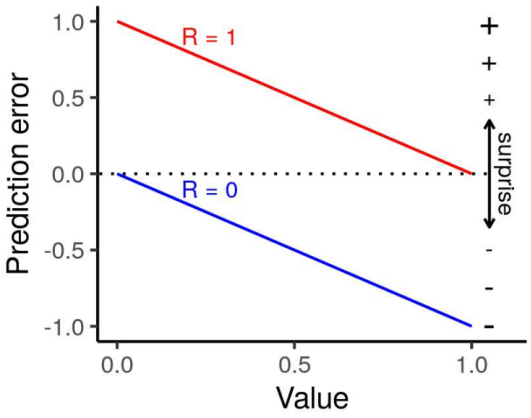

FIGURE 4.B.2: Illustration of the negative association between value and prediction error. Higher value of a choice option is always negatively associated with lower prediction error, irrespective of the outcome valence (positive or negative)

.

4.B Negative association between value and prediction error in the Rescorla-Wagner model

The negative association between PE and V is mathematically intuitive. Because $PE = R\check{}V$, if keeping R constant, higher V will surely be related to lower PE . For an intuitive explanation of this negative association, it is first important to note that PE contains a magnitude and a sign. The magnitude (or absolute value) indicates how surprising the outcome is, and the sign indicates the "direction" of the surprise: it is positive if the outcome is better than expected, and negative if the outcome is worse than expected. Translating this to an experiment scenario, for example, at trial t, a participant with a very high subjective value of some choice option chooses this option. According to the task design, one of two outcomes could be observed: The participant can either receive a positive outcome, then the prediction error will be low, in the sense that it will be a positive number with small magnitude; or, the participant can receive a negative outcome, then the prediction error will also be low, in the sense that it will be a negative number with a large magnitude. Either way, high subjective values will always lead to lower prediction errors ("less positive" or "more negative") than low subjective values (see Figure 4.B.2 for an illustration). The reverse is also true if the participant has a very low subjective value.

4.C Model-based fMRI analysis of reward prediction error

We recommend a two-step procedure to perform model-based fMRI analysis of reward prediction error: first identify, then justify.[1]

First, we constructed and estimated a first-level general linear model (GLM) with parametric modulations (PM) of reward prediction error (PE) at the onset of the outcome, and then tested the significance at the second level. Note that only one parametric regressor of PE was entered into this GLM. This step identified brain area(s) (e.g., nucleus accumbens, NAcc) that were parametrically associated with PE signals derived from the reinforcement learning model. This step is commonly done with neuroimaging analysis tools, like SPM, FSL, etc.

Second, we sought to justify whether the brain activation found in the previous step was indeed associated with PE, rather than outcome valence. We followed the procedure established by previous studies (Behrens et al., 2008; Jocham et al., 2014; Klein et al., 2017) to perform the region of interest (ROI) time series estimates. We first extracted raw blood-oxygen-level dependent (BOLD) time series from the ROI of NAcc. The time series of each participant was then time-locked to the beginning of each trial with a duration of 30 s. All these time points corresponded to the mean onsets for each cue across trials and participants. Afterward, time series were up-sampled to a resolution of 250 ms (1/10 of the repetition time, TR = 2.5 s) using 2D cubic spline interpolation, resulting in a data matrix of size m-by-n, where m is the number of trials, and n is the number of the up-sampled time points (i.e., 30 s / 250 ms = 120 time points). A multiple regression model containing the parametric modulators (e.g., two subcomponents of PE: reward R and value V) was then estimated at each time point (across trials) for each participant. It should be noted that, although the linear regression here took a similar formulation as the first-level general linear model (GLM), it did not model any specific onset; instead, this regression was fitted at each time point in the entire trial across all the trials. Note also that this multiple regression analysis can be done in any statistical software, like Matlab, R, etc. The resulting time courses of effect sizes (regression coefficients) were finally averaged across participants.

It is worth noting that, when the correlation between R and V is noticeable, orthogonalizing V with respect to R might be considered. Although the procedure of the proposed identify-then-justify approach remains unaffected, we recommend identifying possible correlated regressors using simulations even before data collection.

To test group-level significance of the above time series analysis, we employed a permutation procedure. For the time sources of effect sizes for each ROI, we defined a time window of 3–7 s after the corresponding event onset, during which the BOLD response was expected to peak. In this time window, we randomly flipped the signs of the time courses of effect sizes for 5,000 repetitions to generate a null distribution, and tested whether the mean of the generated data from the permutation procedure was smaller or larger than 97.5% of the mean of the empirical data.

These analyses and the visualization were performed with MATLAB R2017a

(https://www.mathworks.com).

[1] This section is adapted from (Zhang & Gläscher, 2019).

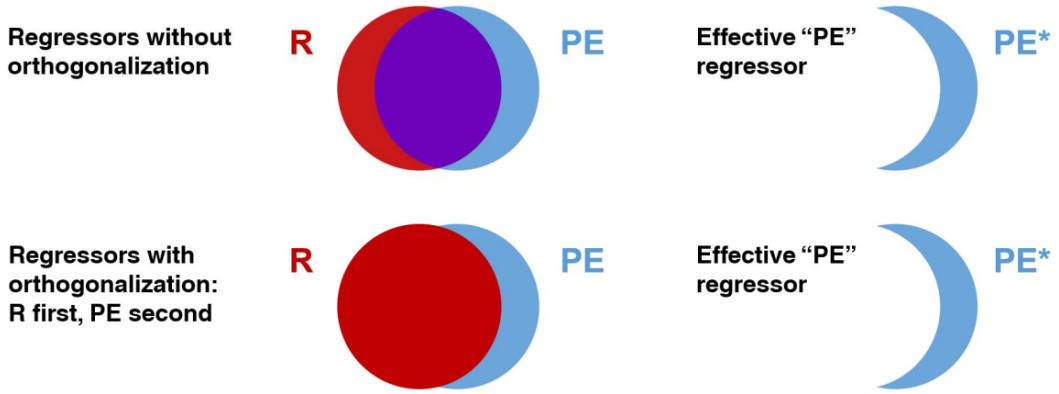

FIGURE 4.D.3: Descriptive variability with two parametric modulators, reward (R) and prediction error (PE). The top row shows how the variability is shared when there is no orthogonalization. The bottom row shows how the variability is distributed when using orthogonalization when R is entered before PE (i.e., PE is orthogonalized with respect to R). Note that in both cases, the effective component of PE that is entered into the model (termed as PE*) is the same. What is crucial to recognize (and often misunderstood) is that, because PE = $R - V$, the PE* here is qualitatively equivalent to $-V$ (negation of V), and brain areas resulted from either of these two approaches might effectively reflect their negative association with value, rather than a positive association with PE.

4.D Discussion of alternative approaches to performing model-based fMRI analysis of reward prediction error

We are aware of at least two alternative approaches reported in previous studies to identify brain regions that are associated with the PE signal (see Figure 4.D.3 for an illustration). The first approach includes both R and PE (without orthogonalization) in the first-level GLM, and the second approach includes PE orthogonalized with respect to R in the first-level GLM. In either of the above approach, although the effective component of PE (termed as PE*) that is entered into the model is the same (Figure 4.D.3; see Mumford et al. 2015 for a discussion), the estimation and the interpretation might be biased. What is crucial to recognize (and often misunderstood) is that, because PE $= R - V$, the PE* here is approximately equivalent to $-V$ (negation of V). We ran a simple simulation, and we found that PE* and $-V$ are highly correlated ($r = 0.8022$ averaged across 1,000 simulations with $\alpha = 0.3$ and $\tau = 3$ using a standard Rescorla-Wagner model). Consequently, brain areas resulted from either of these two approaches might effectively reflect their negative association with value, rather than a positive association with PE. We, therefore, are inclined to discourage both approaches and recommend our approach described in Appendix 4.C.

4.E Posterior predictive check for the simple reinforcement learning model

Choice data were generated using the simple Rescorla-Wagner reinforcement learning (RL) model for 10 virtual participants, with a fixed reward schedule 75:25. The learning rate (α) is sampled from a normal distribution with a mean of 0.6 and a standard deviation of 0.12, and the inverse temperature (τ) is sampled from a normal distribution with a mean of 1.5 and a standard deviation of 0.2. Choice accuracy, namely, how often the more rewarding option is selected, was computed for three levels: (1) trial-level when averaging across participants, (2) participant-level when averaging across trials, and (3) overall level when averaging across both trials and participants.

We then fit the same RL model to the generated data from those 10 virtual participants with hierarchical Bayesian analysis (HBA; Gelman et al. 2013) using a newly developed statistical computing language Stan (Carpenter et al., 2017) in R. Stan utilizes a Markov Chain Monte Carlo (MCMC) sampling scheme to perform full Bayesian inference and obtain the actual posterior distribution. We performed HBA rather than maximum likelihood estimation (MLE) because HBA provides much more stable and accurate estimates than MLE (Ahn et al., 2017). Following the approach in the "hBayesDM" package (Ahn et al., 2017), we assumed, for instance, that a generic individual-level parameter ϕ was drawn from a group-level normal distribution, namely, $\phi \sim \mathcal{N}(\mu_\phi, \sigma_\phi)$, with $\mu\phi$ and $\sigma\phi$ being the group-level mean and standard deviation, respectively. Both these group-level parameters were specified with weakly-informative priors (Gelman et al., 2013): $\mu_\phi \sim \mathcal{N}(0, 1)$ and $\sigma_\phi \sim \text{HalfCauchy}(0, 5)$. This was to ensure that the MCMC sampler traveled over a sufficiently wide range to sample the entire parameter space. In HBA, all group-level parameters and individual-level parameters were simultaneously estimated through the Bayes' rule by incorporating behavioral data. We fit each candidate model with four independent MCMC chains using 1,000 iterations after 1,000 iterations for the initial algorithm warmup per chain, which resulted in 4,000 valid posterior samples. Convergence of the MCMC chains was assessed both visually (from the trace plot) and through the Gelman-Rubin $\hat{R}$ Statistics (Gelman & Rubin, 1992). $\hat{R}$ values of all parameters were close to 1.0 (all smaller than 1.01 in the current study), which indicated adequate convergence.

In the last step, we tested how well the model's posterior prediction was able to reproduce the behavioral patterns of the data (a.k.a., posterior predictive check, PPC). To this end, we applied a post-hoc absolute-fit approach (Steingroever et al., 2014) to generate predictions from the entire posterior MCMC samples. Namely, we let the model take random draws from each participant's joint posterior distribution to generate choices. We iterated this procedure as many times as the number of samples (i.e., 4,000 times) per trial and per participant. By doing so, we tested whether the generated data could capture the behavioral pattern in our behavioral analysis. With these model predicted choices (i.e., 4,000 model generated choices per trial per participant), we computed the choice accuracy the same way as we did before: (1) trial-level when averaging across participants, (2) participant-level when averaging across trials, and (3) overall level when averaging across both trials and participants.

These analyses were performed with R 3.6.1 (R Core Team, 2021). The visualization was performed with MATLAB R2017a (https://www.mathworks.com).

4.F Comparing group differences using hierarchical modeling

We propose two ways to construct hierarchical models for the purpose of group comparison (e.g., comparing clinical vs. non-clinical populations, or treatment vs. control groups) and discuss their application.

The first approach uses two separate hierarchical structures to estimate parameters. Using θ to denote a generic parameter, these separate hierarchical structures for a comparison between a patient group and a control group can be constructed as follows:

$$\begin{aligned}\theta_{\text{patient},i} &\sim \mathcal{N}(\mu_{\theta,\text{patient}}, \sigma_{\theta,\text{patient}}),\\ \theta_{\text{control},i} &\sim \mathcal{N}(\mu_{\theta,\text{control}}, \sigma_{\theta,\text{control}}),\end{aligned}$$

where μ, σ respectively denote the group-level mean and standard deviation, and the index i denotes each individual per group. The second approach includes a higher population level (regardless of patients vs. control) in addition to the previous approach. Using μ_θ and σ_θ to denote the mean and standard deviation of the population-level parameter, this "single group" hierarchical structure is constructed as follows:

$$\begin{aligned}\mu_{\theta,\text{patient}} &\sim \mathcal{N}(\mu_\theta, \sigma_\theta),\\ \mu_{\theta,\text{control}} &\sim \mathcal{N}(\mu_\theta, \sigma_\theta),\\ \theta_{\text{patient},i} &\sim \mathcal{N}(\mu_{\theta,\text{patient}}, \sigma_{\theta,\text{patient}}),\\ \theta_{\text{control},i} &\sim \mathcal{N}(\mu_{\theta,\text{control}}, \sigma_{\theta,\text{control}}).\end{aligned}$$

Employing the separate-group approach or the single-group approach largely depends on the research question in practice. If the goal of the study is to find potentially different cognitive mechanisms between groups, then fitting as separate groups seems more appropriate, because this allows finding distinct winning models per group. By contrast, if the goal is to examine the potential difference in parameters within the "overall" winning model, it is then more reasonable to fit as a single group, with the multi-hierarchy approach described above (see Valton et al. 2020 for a discussion).

Chapter 5

When Implicit Prosociality trumps Selfishness: The Neural Valuation System Underpins more Optimal Choices when Learning to Avoid Harm to Others than to Oneself

Abstract

Humans learn quickly which actions cause them harm. As social beings, we also need to learn to avoid actions that hurt others. It is currently unknown if humans are as good at learning to avoid others' harm (prosocial learning) as they are at learning to avoid self-harm (self-relevant learning). Moreover, it remains unclear how the neural mechanisms of prosocial learning differ from those of self-relevant learning. In this fMRI study, 96 male human participants learned to avoid painful stimuli either for themselves or for another individual. We found that participants performed more optimally when learning for the other than for themselves. Computational modeling revealed that this could be explained by an increased sensitivity to subjective values of choice alternatives during prosocial learning. Increased value-sensitivity was further associated with empathic traits. On the neural level, higher value-sensitivity during prosocial learning was associated with stronger engagement of the ventromedial prefrontal cortex (VMPFC) during valuation. Moreover, the VMPFC exhibited higher connectivity with the right temporoparietal junction during prosocial, compared to self-relevant, choices. Our results suggest that humans are particularly adept at learning to protect others from harm. This ability appears implemented by neural mechanisms overlapping with those supporting self-relevant learning, but with the additional recruitment of structures associated to the social brain. Our findings contrast with recent proposals that humans are egocentrically biased when learning to obtain monetary rewards for self or others. Prosocial tendencies may thus trump egocentric biases in learning when another person's physical integrity is at stake.

This chapter is published as **Lengersdorff, L. L.**, Wagner, I. C., Lockwood, P. L., & Lamm, C. (2020). When implicit prosociality trumps selfishness: the neural valuation system underpins more optimal choices when learning to avoid harm to others than to oneself. *The Journal of Neuroscience*, *40*(38), 7286–7299. https://doi.org/10.1523/JNEUROSCI.0842-20.2020

Significance Statement

We quickly learn to avoid actions that cause us harm. As "social animals," we also need to learn and consider the harmful consequences our actions might have for others. Here, we investigated how learning to protect others from pain (prosocial learning) differs from learning to protect oneself (self-relevant learning). We found that human participants performed better during prosocial learning than during self-relevant learning, as they were more sensitive toward the information they collected when making choices for the other. Prosocial learning recruited similar brain areas as self-relevant learning, but additionally involved parts of the "social brain" that underpin perspective-taking and self-other distinction. Our findings suggest that people show an inherent tendency toward "intuitive" prosociality.

5.1 Introduction

To ensure survival, it is essential that we learn to refrain from actions that cause ourselves harm. Physical pain acts as a powerful learning signal, indicating the immediate need to adjust our behavior to avoid injury (Tabor & Burr, 2019; Vlaeyen, 2015; Wiech & Tracey, 2013). As social beings, we also have to learn and adapt our behavior to avoid harm to others, and "interpersonal harm aversion" has been proposed as the basis of prosocial behavior and morality (C. Chen et al., 2018; Crockett, 2013; Decety & Cowell, 2018; Gray et al., 2012). However, it remains unknown if humans are as good at learning to avoid harm to others (*prosocial learning*) as they are at learning to avoid harm to themselves (*self-relevant learning*). Moreover, while the neural underpinnings of self-relevant learning are well-established, the mechanisms behind prosocial learning remain unclear.

Recent evidence suggests that other people's physical integrity is a highly relevant factor for human behavior. Crockett and colleagues Crockett et al. (2014, 2015) found that individuals are willing to spend more money to protect others from pain than to protect themselves. This suggests that individuals are "hyperaltruistic" in situations where they can deliberately weigh self- and other-relevant outcomes. It remains unclear, though, if hyperaltruism extends to situations where prosociality depends on more implicit processes, such as operant learning (Zaki & Mitchell, 2013). Other studies have proposed an egocentric bias in learning: Participants learned more slowly to gain financial rewards, and to avoid financial losses, for others than for themselves (Kwak et al., 2014; Lockwood et al., 2016), and to associate objects with others compared to oneself (Lockwood et al., 2018). Crucially, it is thus possible that humans show superior self-relevant learning in the context of financial outcomes and basic associative learning, but similar, or superior, prosocial learning when another person's health is at stake. Moreover, prosocial learning performance might vary greatly between individuals, due to differences in socio-cognitive traits such as empathy (Lamm et al., 2019; Lockwood et al., 2016; Olsson et al., 2016). Here, we therefore investigated people's performance in prosocial learning in the context of harm-avoidance, and how it is associated with empathic traits.

Advances in the neuroscience of reinforcement learning (D. Lee et al., 2012) suggest that self-relevant learning is implemented by neural systems for two processes: *valuation*, i.e., choosing between alternatives based on their subjective values, and *outcome evaluation*, i.e., updating values in response to choice outcomes. Converging evidence suggests that both self- and other-relevant valuation engages the ventromedial prefrontal cortex (VMPFC; Kable & Glimcher, 2009; Ruff & Fehr, 2014). Activation of the VMPFC related to valuation affecting others appears further modulated by brain areas connected to social cognition, such as the temporoparietal junction (TPJ),

a region linked to self-other distinction and perspective-taking (Hare et al., 2010; Janowski et al., 2013). With respect to outcome evaluation, it has been found that learning based on pain as a feedback signal engages the anterior cingulate cortex (ACC) and the anterior insula (AI), both when pain is directed to oneself, and when witnessing pain in conspecifics (Keum & Shin, 2019; Lindström et al., 2018; Olsson et al., 2007). It is thus likely that these brain areas also underpin prosocial learning, but the extent of their involvement is unknown.

Here, we conducted a high-powered study with male human participants (N=96) who learned to avoid painful stimuli either for themselves or another individual. Combining computational modeling with fMRI, we tested whether participants were better, or worse, at learning to avoid others' harm compared to self-harm. We expected that differences in learning behavior should be reflected by different recruitment and connectivity of the VMPFC during valuation, as well as the ACC and AI during outcome evaluation. Our major aim was to clarify whether humans are actually "selfish" learners, as suggested by previous evidence using monetary outcomes as learning signals, or if prosocial tendencies can trump egocentricity in the face of harm.

5.2 Materials and Methods

Data reported here were acquired as part of a longitudinal project investigating the effects of violent video games on social behavior (see Chapter 2). In brief, participants completed two fMRI sessions approximately two weeks apart. After the first session, participants came into the behavioral laboratory for seven times over the course of two weeks to complete a violent (or non-violent) video game training. Here, our focus is exclusively on the data from the first session, prior to video game training. Note that at this point of the experiment, participants had not yet been exposed to any experimental manipulation, instructions, or other types of information associated with violent video games. This ensures that the present results are not influenced by the design of the overarching study, and that we could exploit the high sample size of that study to pursue the present scientific aims related to prosocial learning.

5.2.1 Participants

Ninety-six male volunteers (age range: 18 - 35 years) participated in the study. The novelty of the present design precluded formal power analysis. As a benchmark, though, we considered the effect size reported by Lockwood et al. (2016), who used a similar task, but with financial rewards as positive reinforcers, rather than pain as negative reinforcers. They found an effect size of d = 0.87 when comparing the learning rates for self-relevant learning to the learning rates for prosocial learning. With a sample size of 96, we had a power > 99.9% to find an effect of this size in our study. Sensitivity analysis further indicated that we had a power of 80% to detect an effect of d = 0.29, suggesting that our study was adequately powered to detect medium-to-small effects (J. Cohen, 2013). Power analyses were performed with the software *Gpower 3* (Faul et al., 2007). All participants were healthy, right-handed, had normal or corrected-to-normal vision, reported no history of neurological or psychiatric disorders or drug abuse, and fulfilled standard inclusion criteria for MRI measurements. Only male participants of the age range 18 – 35 years were tested based on constraints related to the overarching project on violent video games. Hence, the findings presented here only apply to this population. All participants

gave written consent prior to participation and received financial reimbursement, and the study had been approved by the Ethics Committee of the Medical University of Vienna.

5.2.2 Study timeline and procedures

Participant arrival and interaction with the confederate

Each participant was paired with another male participant who in reality was a member of the experimental team (i.e., a confederate). This way, we ensured that the participant was under the impression of completing the prosocial learning task together with another person (see below for details of the task). As a cover story, the participant was told that he had been randomly assigned to the role of the "MRI participant", and would complete the experimental tasks inside the MR scanner, while the confederate had been assigned to the role of "pulse measures participant" from whom we would take pulse measurements only. After arrival at the MR facility, it was explained that the participant would first receive task explanations in another room while the confederate would complete the pain calibration, and that they would switch places afterwards (only the participant actually underwent the experimental procedures). After the pain calibration (see below), the participant was positioned in the MR scanner and the confederate was seated next to the scanner at a table with an MR compatible monitor and keyboard. To uphold the impression of the confederate actually participating in the task, the confederate remained seated next to the scanner for the entire duration of the experiment, and also responded audibly to the experimenter's mock questions about his physical comfort via the intercom.

Participants were formally interviewed if they doubted the deception (e.g., if they believed that the confederate was a real participant) at the end of the second session of the overarching project (two weeks later). At this point, six participants reported doubts that the confederate was a real participant. No participant expressed doubts about the confederate after the first session, but please note that they were not explicitly interviewed regarding doubts at that point in time, to not raise suspicions about the cover story.

Electrical stimulation and pain calibration

Electrical stimulation was delivered with a Digitimer DS5 Isolated Bipolar Constant Current Stimulator (Digitimer, Ltd., Clinical & Biomedical Research Instruments) using concentric surface electrodes with 7 mm diameter and a platinum pin (WASP electrode, Specialty Developments) attached to the dorsum of the left hand. With this setup, electrical stimulation was constrained to the skin under and directly adjacent to the electrodes. To determine each participant's subjective pain threshold, we employed an adaptive staircase procedure (as previously used in Rütgen, Seidel, Riečanský, & Lamm, 2015; Rütgen, Seidel, Silani, et al., 2015). In brief, the participant received short stimuli (500 ms) of increasing intensity and was asked to rate pain intensity on a ten-point numeric scale from 0 to 9. The numbers corresponded to the following subjective perceptions: 0 = "not perceptible"; 1 = "perceptible, but not painful", 3 = "a little painful", 5 = "moderately painful", 7 = "very painful", 9 = "extremely painful, highest tolerable pain". To account for habituation and familiarization effects, this procedure was repeated once, after which the participant received 30 random electrical stimuli. The average intensities of electrical stimuli that

were rated as 1 and 7 were then chosen as the intensities of non-painful and painful stimulation during the pain-avoidance task.

Prosocial learning task

The participant was instructed to learn the association between abstract symbols and painful electrical stimuli. During each trial (Figure 5.1), the participant had to choose one of two symbols, with one the symbols resulting in non-painful electrical stimulation in 70% of trials, and the other in painful electrical stimulation in 30% of the trials (note that participants were not informed about these contingencies). Importantly, electrical stimulation as a consequence of the choice made was either delivered to the participant (*Self* condition) or to the other person, i.e., the confederate (*Other* condition). For both conditions, the participant completed 3 blocks of 16 trials (resulting in 48 trials per condition, and 96 trials in total). The two experimental conditions alternated per block (so the participants never completed the same condition twice in a row), and the starting condition was counter-balanced across participants. Each block contained a different pair of symbols, and the participant was instructed to learn this new set of symbol-stimulus contingencies once more.

At the start of a new block, the participant was instructed that he would now have to perform the task for himself ("Play for yourself") or for the confederate (e.g., "Play for Michael"). During each trial, the two symbols were presented for 3000 ms and the participant was asked to choose one symbol (choice phase). Choices were made with the right hand via a button box. If the participant made a choice within this time limit, a red or blue arrow (presented for 1000 ms) indicated whether the current recipient would receive painful or non-painful stimulation, respectively (outcome phase). The arrow pointed downwards if the participant received electrical stimulation himself (self-directed electrical stimulation) or to the right towards the confederate (other-directed electrical stimulation). If no button press occurred within the time limit, the message "too slow!" was displayed and was followed by painful electrical stimulation to either the participant or the confederate, depending on the current condition. After an interval of 2500 to 4500 ms (uniformly jittered, in steps of 250 ms), the electrical stimulation was delivered (stimulation phase). During self-directed electrical stimulation (500 ms), the participant saw a pixelated photograph of himself (1000 ms, same onset as electrical stimulus). Pain intensity was indicated by a red or blue lightning icon in the lower right of the photograph. During other-directed electrical stimulation, the participant saw a photograph of the confederate with a neutral or painful facial expression, during non-painful or painful electrical stimulation respectively. The next trial started after an inter-trial interval between 2500 and 4500 ms (uniformly jittered, in steps of 250 ms). The task was presented using COGENT (http://www.vislab.ucl.ac.uk/cogent.php), implemented in MATLAB 2017b (The MathWorks Inc., Natick, MA, USA). The total task duration was approx. 23 min.

As part of the cover story, the participant was instructed that the other participant (i.e., the confederate) would never make choices that could result in stimulation for himself or the participant. Instead, he would be presented with the choices of the participant, and would have to indicate per button press if he would have made the same decision. We chose this approach (instead of saying that the confederate would only passively receive stimulation without any further instructions) to increase the believability of the deception.

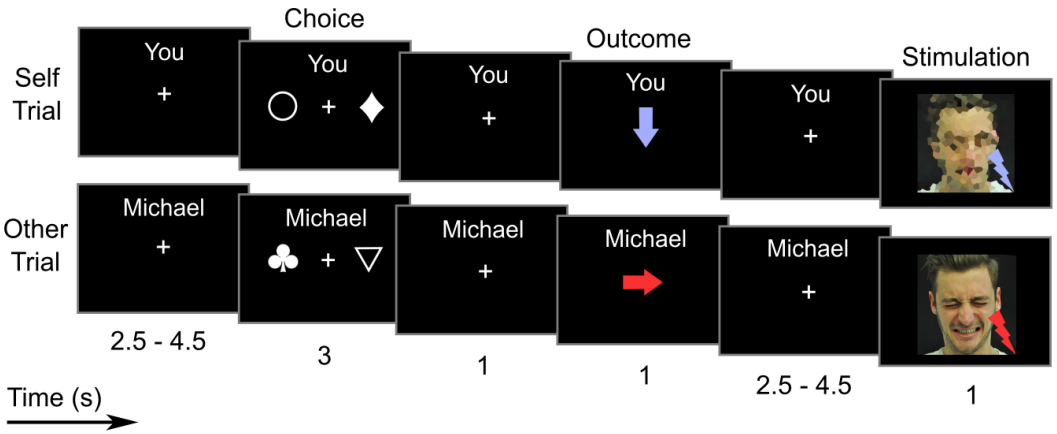

FIGURE 5.1: Schematic depiction of the trial structure of the prosocial learning task. The task was to identify, in the Choice window, one out of two abstract symbols that would prevent painful stimulation from being delivered to either Self (self-relevant learning) or Other (prosocial learning). One symbol had a 30% chance of delivering pain, the other a 70% chance. Feedback on the outcome of the choice was delivered in the Outcome window, by means of color-coded arrows (red: pain; blue: no pain). Overall 16 such trials were performed in each block, with three blocks per condition (*Self* / *Other*; presented in alternating order) being run over the course of the experiment. Upper row: trial of the *Self* condition (self-relevant learning) in which a non-painful electrical stimulus is delivered to the participant as a consequence of his choice. Lower row: trial of the *Other* condition (prosocial learning) in which a painful electrical stimulus is delivered to the confederate as a consequence of the participant's choice. Numbers below descriptions describe event length in seconds, with intervals indicating ranges of interstimulus jitter.

5.2.3 MRI data acquisition

MRI data were acquired with a 3 Tesla Siemens Skyra MRI system (Siemens Medical, Erlangen, Germany) and a 32-channel head coil. Blood oxygen level dependent (BOLD) functional imaging was performed using a multiband accelerated echoplanar imaging sequence with the following parameters: Echo time (TE): 34 ms; Repetition time (TR): 1200ms; flip angle: 66°; interleaved ascending acquisition; 52 axial slices coplanar to the connecting line between anterior and posterior commissure; multi-band acceleration factor 4, resulting in 13 excitations per TR; field-of-view: 192 ×192 × 124.8 mm, matrix size: 96 × 96, voxel size: 2 × 2 × 2 mm, interslice gap 0.4 mm. Structural images were acquired using a magnetization-prepared rapid gradient-echo sequence with the following parameters: TE = 2.43 ms; TR = 2300 ms; 208 sagittal slices; field-of-view: 256 × 256 × 166 mm; voxel size: 0.8 × 0.8 × 0.8 mm. To correct functional images for inhomogeneities of the magnetic field, field map images were acquired using a double echo gradient echo sequence with the following parameters: TE1/TE2: 4.92/7.38 ms; TR = 400 ms; flip angle: 60°; 36 axial slices with the same orientation as the functional images; field-of-view: 220 × 220 × 138 mm; matrix size: 128 × 128 × 36; voxel size: 1.72 × 1.72 × 3.85 mm.

5.2.4 Trait measures

As a measure of trait empathy, participants completed the Questionnaire of Cognitive and Affective Empathy (QCAE) (Reniers et al., 2011), which assesses empathic traits along the two dimensions *cognitive empathy* (with the subdimensions *perspective taking* and *online simulation*) and *affective empathy* (with the subdimensions *emotion contagion*, *proximal responsivity* and *peripheral responsivity*). The questionnaire was completed after the MR session. Due to an error in data collection, questionnaire data from nine participants were not available for analysis. We therefore restricted analyses related to trait measures of empathy to the remaining 87 participants.

5.2.5 Data analysis

Our analysis plan was as follows: first, we tested whether participants showed better learning behavior during prosocial or self-relevant learning, as measured by an objective index of optimal behavior (described in detail below). Then, we used computational modeling to investigate if behavioral differences could be explained by differences in the mechanisms underlying these two types of learning. Finally, we used fMRI analyses to relate differences on the behavioral level to differences in the neural mechanisms.

Analysis of optimal behavior

We first investigated if participants showed more optimal behavior during prosocial or self-relevant learning. In decision theory, *optimal choices* are defined as those choices that maximize the expected rewards (or minimize the expected loss), given all available information (Kulkarni & Gilbert, 2011). Mathematically, our learning task is equivalent to a two-armed bandit problem with unknown reward probabilities of the two choice alternatives, p_A and p_B. For this kind of problem, the optimal choice in each trial can be derived algorithmically, as described by Steyvers et al. (2009) (see also Kaelbling et al., 1996). For mathematical details, we refer the reader to these references, and to our commented scripts (see section **Code and data availability**). In short, a choice is optimal if an "ideal participant" with perfect memory and the

ability to calculate all possible outcome sequences would make the same choice, given the information at hand. Importantly, an optimal choice maximizes not only the expected outcome of the current trial, but also the possible outcomes of subsequent trials, balancing exploration and exploitation. Optimal choices also depend on the prior information the decider has about the choice alternatives. This can be modelled by defining a prior distribution for the reward probabilities p_A and p_B. Following Steyvers et al. (2009), we modelled prior information about the learning task with a Beta distribution, with parameters a and b set to 1.5. This gives a symmetric distribution with considerable probability mass for all but the most extreme values of p_A and p_B. Thus, the distribution reflects that participants were informed that both symbols will lead to painful stimulation in some cases, but did not get any further information about the probabilities. To investigate how sensitive our results were to these exact values of a and b, we also calculated optimal choices under Beta distributions with both values set to 1 (reflecting a uniform distribution) or 2 (putting more probability mass on average probabilities).

Preprocessing, analysis and visualization of behavioral data were performed using R 3.4.4 (R Core Team, 2021). To test for differences in the number of optimal choices between prosocial and self-relevant learning, we computed a generalized linear mixed model (GLMM), with optimal choice as binary response variable (binomial family, logit link function), using the R packages `lme4` (Bates et al., 2015) and `afex` (Singmann et al., 2020). *Trial* (1 to 16; numeric, mean centered), *condition* (*Self* vs. *Other*; categorical dichotomous; coded as Self = -1, Other = 1) and their interaction *trial x condition* were specified as predictors. To account for dependencies between observations within the same participants, we treated participants as levels of a random factor. Following the recommendation of (Matuschek et al., 2017), we selected the random effects structure that led to the most parsimonious fit of the GLMM, as indicated by the Akaike Information Criterion (AIC; Akaike, 1998) We then tested the significance of the fixed effects of the final model, using Type III drop-one-term likelihood ratio tests, as implemented by the `mixed` function of the package `afex`.

Computational modeling

We used computational modeling to investigate differences in the processes underlying prosocial and self-relevant learning. In computational modeling, one first defines a model as a set of mathematical equations that relate observable behavior to theoretical quantities. If the model is true (or, at least, useful; Box, 1976), it should help explain the observed behavior of a participant. Importantly, the exact relationship between a models' theoretical quantities and actual behavior is governed by a number of free parameters, which have to be estimated from the data. Using techniques of model comparison, one can then select the most useful model from a set of plausible candidates. Here, we utilized reinforcement learning (RL) models, which already have a long history of use in neuroscientific research (Hackel & Amodio, 2018). Following this modeling framework, we assumed that participants chose between symbols A and B based on the symbols' subjective values (valuation). Then, they updated these subjective values based on the outcome of their choice (outcome evaluation).

Models for valuation We used the *softmax function* to model how subjective values are mapped onto choice probabilities. This function takes as input the subjective values of symbols A and B (here denoted as V_A and V_B, and bound in the interval

[0,1]), and gives as output the probability of choosing one symbol over the other. In the case of only two alternatives/symbols, the softmax simplifies to the logistic function, and is thus

$$P\left(choice_t = A|V_{A,t}, V_{B,t}\right) = \frac{1}{1 + e^{-\beta(V_{A,t} - V_{B,t})}},$$

where $P\left(choice_t = A|V_{A,t}, V_{B,t}\right)$ is the probability of choosing symbol A in trial t, given specific values of $V_{A,t}$ and $V_{B,t}$, e is the basis of the exponential function, and β is the free inverse temperature parameter, defined on the positive real numbers (0, ∞). Note that in the simple case of two symbols, the probability of choosing symbol "B" is given by $1 - P\left(choice_t = A|V_{A,t}, V_{B,t}\right)$. The probability of choosing symbol A increases nonlinearly (in a sigmoid shape) with the difference $V_{A,t} - V_{B,t}$; for large positive values of the difference, the probability approaches 1, while for strongly negative values of the difference, the probability approaches 0. If the difference is close to 0, the probability is close to 0.5, that is, the choices of the participant are nearly random. Importantly, the free parameter β defines how sensitive the softmax function, and thus the participants' choices, are to the difference in subjective values. For very high β, even small differences in subjective values lead to a high probability of choosing the symbol with the higher value. For β close to 0, choices are very random even if the difference in subjective values is high. In this paper, we will refer to β as a measure of value-sensitivity, consistent with the parameter's mathematical definition in the RL model (see also Chung et al., 2017; Katahira, 2015). Importantly, this interpretation is also consistent with the interpretation of β as an indicator of the exploration/exploitation trade-off (e.g., Cinotti et al., 2019; Humphries et al., 2012). In other words, a high β may indicate exploitative choices, as choice behavior is very sensitive to differences in value, and the option with the higher value is frequently chosen. In contrast, a low β leads to choices that are less dependent on the value difference, which may reflect the exploration of the alternative with the lower values.

Importantly, to test for differences between self-relevant and prosocial learning, we also considered models with condition-specific parameters (see Lockwood et al., 2016, for a similar approach). Thus, we also estimated models in which different inverse temperature parameters β_{Self} and β_{Other} governed participants' choice behavior during self-relevant and prosocial learning, respectively. Differences here would indicate that a participant showed differences in value sensitivity between the two conditions, i.e., for self-relevant and prosocial learning.

Models for outcome evaluation Following the RL modeling framework, we assumed that participants updated the subjective value of the chosen symbol based on the received outcome. Here, we will refer to outcomes that indicate the delivery of non-painful stimulation as *positive* outcomes, and to outcomes that indicate painful stimulation as *negative* outcomes. According to the simplest RL model (Hackel & Amodio, 2018; Sutton & Barto, 2018), subjective values are updated according to the formula

$$V_{A,t+1} = V_{A,t} + \alpha\left(R_t - V_{A,t}\right),$$

where $V_{A,t}$ is the value of the chosen symbol A in trial t, $V_{A,t+1}$ is the updated value, and α is the free learning rate parameter, which is bound in the interval [0,1]. R_t is the

number encoding the outcome of trial t, here defined as 1 for positive outcomes/"no pain", and 0 for negative outcomes/"pain". We chose this coding scheme to make our algorithm more comparable to those used in other studies using the RL model (e.g., Lockwood et al., 2016). But note that this coding scheme is mathematically equivalent to that used in other aversive learning paradigms where pain outcomes are instead coded as -1 and avoidance of pain as 0 (e.g., Roy et al., 2014; Seymour et al., 2004, for a mathematical proof of this statement, see https://osf.io/h9txe/). The term $(R_t - V_{A,t})$ is usually referred to as the prediction error PE_t of trial t. The free parameter α controls the degree to which the value of the chosen symbol is updated by the prediction error. For α close to 0, there is nearly no updating. For α close to 1, the subjective value is strongly updated by the last outcome. Note though that the learning rate α cannot directly be interpreted as a measure of how fast participants understand which of the two symbols is better (see also Zhang et al., 2020). Rather, it describes how strongly subjective values are influenced by the last outcome, regardless of the outcomes that came before. $V_{A,0}$ and $V_{B,0}$ were both initialized to 0.5, reflecting the assumption that participants were equally likely to prefer either symbol on the first trial, and that they had no prior information about symbol-outcome contingencies.

Participants might have responded differently to positive outcomes (i.e., non-painful stimulation) than to negative outcomes (i.e., painful stimulation). We therefore also considered an extension of the simple RL model which allowed for outcome-specific learning rate parameters (Den Ouden et al., 2013). In this model, participants weighted positive prediction errors with the learning rate α^+, and negative prediction errors with the learning rate α^-. Crucially to our research question, we also estimated models in which condition-specific learning rate parameters α_{Self} and α_{Other} controlled the rate with which participants' updated the subjective values during self-relevant and prosocial learning, respectively. If we found beforehand that outcome-specific learning rate parameters are of importance, we also tested models with two separate learning rates for positive outcomes ($\alpha^+_{Self}$and α^+_{Other}), two separate learning rates for negative outcomes (α^-_{Self} and α^-_{Other}), or both.

Hierarchical modeling

The computational models described so far explain a single participant's behavior. To combine the data from all participants in our sample, and draw inferences about the underlying population, we used hierarchical Bayesian modeling. We thus modelled how the parameters that characterized participants' behavior were distributed in the population. Modeling the commonalities between participants in such a way has been shown to enable more stable individual parameter estimates and allows direct inference about population parameters (Ahn et al., 2017).

Individual learning rate parameters were modelled to follow logit-normal distributions with unknown population mean μ_α and standard deviation σ_α. This defines a normal distribution on the unbounded scale of the logits of the learning parameters, which are then mapped into the interval [0,1] by means of the logistic function. Inverse temperature parameters were modelled to follow log-normal distributions with unknown population mean μ_β and standard deviation σ_β. This defines a normal distribution on the unbounded scale of the natural logarithms of the parameters, which are then sent to the positive real numbers with the exponential function.

Model estimation

Bayesian model estimation was performed using the software STAN v.2.18.1 (Carpenter et al., 2017) implemented in R. For a detailed description of the hierarchical models, see Appendix 5.A. Custom code can be found online (https://osf.io/53qvd/). STAN utilizes Markov Chain Monte Carlo (MCMC) sampling to approximate posterior distributions of parameters in light of observed data. MCMC convergence diagnostics were performed by visual inspection of the trace-plots and standard diagnostics output of the software (Gelman & Rubin, 1992). We performed model comparison and selection by treating the models to be compared as components of an overarching mixture model (Kamary et al., 2014; Keller & Kamary, 2017; Ly et al., 2016; Robert, 2016). This approach allows to assess which of the compared models is most supported by the observed data, through the calculation of Bayes factors (BF; Kass & Raftery, 1995). We used the following rationale to find the best-fitting model: We first assessed whether the data were better explained by a model containing outcome-specific learning rate parameters (α^+, α^-). We then assessed in a step-wise manner if the model fit increased by adding separate parameters for self-relevant and prosocial learning. Here, we considered both fixed effects (systematic differences in parameters that characterize the whole population) as well as random effects (random subject-wise differences between parameters). To test whether the addition of a condition-wise parameter difference increased the model fit, we calculated a mixture model containing the current model without the additional parameter, a model including the respective random effect, and a model including the random effect as well as the fixed effect. The mixture model was estimated with 8 MCMC chains of 1000 warm-up samples, and 4000 actual samples. We only added a parameter difference to the model if the model including the parameter difference had a substantially better model fit compared to the model without the parameter difference, as indicated by a Bayes factor > 3 (Kass & Raftery, 1995).

After model selection, we re-estimated the winning model with 4 chains of 2000 warm-up samples and 5000 actual samples. Single-value individual parameter estimates were calculated as the posterior means of the parameters. To assess absolute, rather than relative, model validity, we performed posterior predictive checks (PPC; Gelman et al., 2013; Zhang et al., 2020). For each draw of the posterior distribution of parameters, we simulated the behavior of new participants with the same set of parameters. For visual PPC, we plotted the 95% highest-density intervals of participants' predicted choices against their actual choices. Furthermore, we calculated the correlation coefficients between participants' actual percentage of optimal choices, and the mean of the predicted percentage of optimal choices.

Associations between model parameters and optimal choices

To formally test if differences in optimal choice behavior could be explained by differences in learning mechanisms, we conducted multilevel mediation analysis (Bauer et al., 2006; Kenny et al., 2003). We defined *condition* (Self vs. Other) as independent variable, relevant model parameters as mediator, and optimal choice as dependent variable. We estimated the effect of *condition* on the mediator (path a) with linear mixed regression, using the function "lmer" of the package `lme4` (Bates et al., 2015). We estimated the effect of the mediator on optimal choices (path b) by adding the mediator to the GLMM described above (section *Analysis of optimal behavior*). To make inference about the indirect effect of *condition* via the mediator, we calculated bias-corrected and accelerated confidence intervals (BCa-CI; Efron

1987) for the product *a*b*, using 1000 bootstrapping samples with the R package `boot` (Davison & Hinkley, 1997), and tested if this interval contained zero. BCa-CIs correct for bias and skewness of the bootstrap sample distribution.

Associations between prosocial behavior and trait empathy

Using regression analysis, we investigated whether differences in prosocial versus self-relevant learning behavior were associated with empathic traits. We regressed the difference in parameters for prosocial versus self-relevant learning on the subscales of the QCAE. To test whether the analysis suffered from multicollinearity due to potentially high correlations between the subscales, we also calculated the variance inflation factors per predictor.

fMRI data preprocessing

Preprocessing and analysis of fMRI data were performed using SPM12 (Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm) implemented in MATLAB 2017b. Functional images were slice timed and referenced to the middle slice, realigned to the mean image, and unwarped to correct for motion-by-distortion interaction artifacts using the acquired field map. The structural image was then co-registered to the mean image of the realigned functional images using mutual information maximization, and structural and functional images were normalized to the stereotactic Montreal Neurological Institute (MNI) space (template: MNI ICBM152). The normalized functional images were smoothed with a Gaussian kernel (4 mm, full-width-at-half-maximum). To remove motion-related artifacts, the functional images were then subjected to an independent-component-analysis based algorithm for automatic removal of motion artifacts (ICA-AROMA; Pruim, Mennes, Buitelaar, & Beckmann, 2015; Pruim, Mennes, van Rooij, et al., 2015), implemented using the FMRIB software library (FSL v5.0; http://www.fmrib.ox.ac.uk/fsl). For each experimental run, the time course was decomposed into 100 independent components (ICs) which were classified as motion- or signal-related ICs. Following this, motion-related ICs were regressed from the time course using ordinary least-squares regression. To further identify participants who exhibited extreme head movement during experimental runs, we calculated the framewise displacement (FD; Power et al., 2012) as a measure of relative image-to-image motion. Since we *a priori* had decided to exclude runs of participants who exhibited FD > 0.2 (indicating voxel displacement of approximately 2 mm between volumes) in more than 1% of volumes from all further analyses, data of five participants were excluded from the analyses. Data moreover was high-pass filtered with the SPM12 standard cutoff of 128 s.

fMRI data analysis

For the prosocial learning task, first-level design matrices of the general linear model (GLM) included regressors for the choice phase (presentation of the symbols, 3000 ms), the outcome phase (presentation of the arrow cues indicating the outcome, 1000 ms), and the stimulation phase (electrical stimulation and presentation of the photographs, 1000 ms; see also Figure 5.1). Regressors were built by creating onset-locked boxcar-functions with the respective event durations, convolved with the canonical hemodynamic response function. For the choice phase, regressors were defined separately for the *Self* and the *Other* condition. For the outcome phase, regressors were defined separately for the *Self* and the *Other* condition, as well

as for *positive* outcomes (indicating the delivery of non-painful stimulation) and *negative* outcomes (indicating the delivery of painful stimulation). For the stimulation phase, regressors were created separately for every combination of the factors *Self* vs. *Other* and *non-painful* stimulation (after positive outcomes) vs. *painful stimulation* (after negative outcomes). Missed responses were collapsed within three additional nuisance regressors during choice, outcome, and stimulation in the condition in which responses were missing. Thus, if a participant had missing responses in both conditions (*Self* and *Other*), six nuisance regressors were added (three per condition, defined as above).

To detect brain areas in which activation was correlated with model-derived quantities, we extracted the values from the winning computational model (see above) and entered them as trial-by-trial parametric modulators. For events of the choice phase, we added the value difference (i.e., value of chosen symbol minus value of unchosen symbol) as a parametric modulator, separately for the *Self* and *Other* conditions. For events of the outcome phase, we added the absolute trial-wise prediction error as a parametric modulator, separately for positive and negative outcomes, and separately for the Self and Other conditions.

We chose a GLM model with separate regressors for positive and negative outcomes, and prediction errors for these outcomes, for two reasons. First, as outcomes and prediction errors are by definition very highly correlated (Behrens et al., 2008; Zhang et al., 2020), this model allowed us to investigate brain activity that is specifically related to deviations from the observed outcome to the expected outcome (note that the parametric modulators for prediction error are orthogonalized to the regressors for outcome, and can therefore only explain that part of signal variance that is not already explained by mere outcomes). Second, we defined different parametric modulators for prediction errors from positive and negative outcomes to be able to investigate where in the brain positive deviations from the expected outcome are processed differently than negative deviations. We used the absolute prediction error to facilitate interpretation of the resulting estimates: prediction errors for positive outcomes were positive, with higher numbers indicating greater deviation from the expectation. Prediction errors for negative outcomes were negative, with lower (more negative) numbers indicating greater deviation from the expectation. Therefore, higher values of the absolute prediction error indicated stronger deviations from the expectation, regardless of the valence of the outcome.

To summarize, the first-level model included 10 task-based regressors, 6 parametric modulators and 0, 3, or 6 nuisance regressors, depending on the number of missed responses per participant. Second-level statistical inference was performed using flexible factorial ANOVAs, one-sample *t*-tests, and paired-sample *t*-tests. The statistical threshold for all whole-brain analyses was defined as $p < 0.05$ family-wise error (FWE) corrected for multiple comparisons at the cluster-level, using a cluster-defining threshold of $p < 0.001$. The corrected cluster size threshold (i.e., the spatial extent of a cluster that is required to be labeled significant) was calculated using the SPM extension `CorrClusTh.m` (script provided by Thomas Nichols, University of Warwick, United Kingdom, and Marko Wilke, University of Tübingen, Germany; http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/scripts/spm/). Anatomical labeling of activation peaks was based on the Automatic Anatomical Labeling atlas (AAL; Tzourio-Mazoyer et al., 2002) and performed using the toolbox `xjView` (http://www.alivelearn.net/xjview).

Functional connectivity analysis

We investigated differences in functional connectivity of the VMPFC during the choice phase, by calculating generalized Psychophysiological Interaction (gPPI) analyses (McLaren et al., 2012). We created an anatomical mask of the VMPFC (taken from the AAL atlas; Tzourio-Mazoyer et al., 2002) with significant clusters that emerged from the activation analysis (correlation of brain activity with the difference in value of chosen symbol minus unchosen symbol). From this intersected mask, the first eigenvariate of the participant-specific functional time course was extracted, adjusted for average activation using an *F*-contrast, and deconvolved to estimate the neural activity in the seed region (i.e., the physiological factor). The estimated neural activity was then multiplied with the boxcar function defining the event of interest (i.e., the psychological factor), and the product was convolved with the hemodynamic response function. This resulted in one psychophysiological interaction regressor per event of interest. The interaction regressors were added to the first-level design matrix and the GLMs were estimated. Group-level connectivity differences between *Other* and *Self* were tested using a paired-sample *t*-test.

5.2.6 Code and data accessibility

Custom scripts for the calculation of optimal choices and estimation of Bayesian models in STAN, as well as a documentation of the used priors, are freely accessible on OSF: https://osf.io/53qvd/. Second-level fMRI contrast maps are accessible at the same location.

5.3 Results

5.3.1 Analysis of optimal choice behavior

We first investigated if participants showed more optimal behavior during self-relevant or prosocial learning, using GLMM. The best fitting GLMM was obtained with a random effects structure containing by-subject random intercepts and random slopes for *condition* (Self vs. Other), *trial (1 to 16)*, and their interaction, see 5.1. This implies that there was substantial variation between participants with respect to the average probability of an optimal choice, how much this probability increased per trial, and how much it differed between conditions. We found that participants made optimal choices significantly above chance level (Intercept = 1.351, $\chi^2_{df=1}$= 93.41, $p < .001$): on average, participants chose optimally in 79.4% of trials (see **Figure 2A**). A significant main effect of *condition* (β = 0.099, $\chi^2_{df=1}$= 5.78, p = .016) indicated that participants made more optimal choices when choosing for the other person compared to when choosing for themselves. Furthermore, there was a significant effect of trial (β = 0.108, $\chi^2_{df=1}$= 55.92, $p < .001$), indicating that the probability of making an optimal choice increased over time. The absence of a significant interaction between condition and trial (β = 0.011, $\chi^2_{df=1}$= 1.62, p = .203) indicated that the increase of optimal choices over time was similar for self-relevant and prosocial learning.

As described in the **Methods** section, we identified optimal decisions under the assumption that the participants' prior knowledge about outcome probabilities could be modelled as Beta(a = 1.5, b = 1.5). To test whether our results were sensitive to this assumption, we repeated the analysis for different prior distributions, Beta(a = 1, b = 1) and Beta(a = 2, b = 2). These analyses led to the same results, indicating that our

TABLE 5.1: Results of the GLMM analysis. Dependent variable: optimal choice (0 = no optimal choice; 1 = optimal choice). Factor coding for Condition: Self = -1; Other = 1. Trial was mean-centered. The column Random effects depicts standard deviations of the random effect associated with the respective model term. p-values were derived from Type-III likelihood ratio tests.

Fixed effects	**Parameter (SE)**	**Random effect**	$\chi^2_{df=1}$	**p-Value**
Intercept	1.351 (0.109)	1.017	93.409	<.001
Condition	0.099 (0.041)	0.222	5.781	0.016
Trial	0.108 (0.013)	0.105	55.919	<.001
Condition × Trial	0.011 (0.008)	0.031	1.623	0.202

findings are robust to changes in prior assumptions (effect of *condition* for Beta(1,1): p = 0.048; for Beta(2,2): p = 0.024).

All participants were able to make a choice within the time limit in the majority of trials. There was no participant who missed more than 10 trials (out of 48) in either condition (number of trials missed in Self condition: 0: 63.54%; 1: 21.87%; 2: 6.25%; 3-10: 8.33%; number of trials missed in Other condition: 0: 57.29%; 1: 25.00%; 2: 6.25%; 3-10: 11.46%). An exploratory GLMM on the number of missed trials revealed no significant difference in probability of missing a trial between the two conditions.

In summary, we found that participants were able to choose optimally between symbols that differed in the probability of delivering painful electrical stimulation. Intriguingly, participants made more optimal choices when choosing for the other person, than when choosing for themselves.

5.3.2 Computational modeling of self-relevant and prosocial learning

Having found that participants made more optimal choices during prosocial than self-relevant learning, we next used computational modeling to investigate if this result could be explained by differences in learning mechanisms. As the first step in model selection, we assessed if the data were better explained by a model with only one learning rate for both positive and negative outcomes (M0), or a model with outcome-specific learning rates (M1). We found that model M1 explained the data substantially better (*BF* M1 vs M0 > 1000). We next compared model M1 against several models with condition-wise differences in a single parameter: a model with different learning rates for positive outcomes α^+ per condition (random effect: M2.1; random effect and fixed effect: M2.2); a model with different learning rates for negative outcomes α^- per condition (random effect: M3.1; random effect and fixed effect: M3.2); and a model with different inverse temperature parameters β per condition (random effect: M4.1; random effect and fixed effect: M4.2). Including different inverse temperature parameters for self-relevant and prosocial learning led to the greatest increase in model fit (*BF* M4.1 vs. M1 = 37.67; M4.2 vs. M1 = 294.67), compared to adding differences in the other parameters (learning rate for positive outcomes: *BF* M2.1 vs. M1 = 8.88; *BF* M2.2 vs. M1 = 9.35; learning rate for negative outcomes: *BF* M3.1 vs. M1 = 0.32; *BF* M3.2 vs M1 = 0.25). We next compared model M4.2 to a model that additionally included different learning rates for positive outcomes (additionally to the differences in inverse temperature; random effect: M5.1; random effect and fixed effect: M5.2). This revealed that model fit was only marginally improved by including the random effect, but not the fixed effect (*BF* M5.1 vs. M4.2 = 1.34; M5.2 vs

M4.2 = 0.92). Finally, we compared model M4.2 to a model that additionally included different learning rates for negative outcomes (M6.1; random effect and fixed effect: M6.2). This revealed that the data was better explained by a model without these additional differences in parameters (*BF* M6.1 vs. M4.2 = 0.28; M6.2 vs M4.2 = 0.23).

Based on this model selection procedure, we selected model M4.2 as the winning model. This model included different learning rates for positive and negative outcomes, and a population-level difference between inverse temperature parameters for self-relevant and prosocial learning. Next, we validated this model using posterior predictive check analyses. Figure 5.2A depicts the 95% highest-density interval (HDI) of the predicted percentage of participants' optimal choices. Participants' behavior appears to be well explained by the model, as actually observed percentages lie well within the HDI. However, note that for trial 3, the percentage of optimal choices was slightly overestimated by the model. Figure 5.2B plots the real percentages of optimal choices against the mean predicted percentages. This plot also indicates suitable model fit, although the model appears to slightly overestimate the number of optimal choices for those participants who made optimal choices below the chance level of 50%. The correlations between the percentage between actually observed choices and the percentage predicted by the model were as follows: optimal decisions for self: mean r = 0.757, 95% HDI = [0.678, 0.831]; optimal decisions for other: mean r = 0.796, 95% HDI = [0.721, 0.867].

The group-level parameter estimates of the winning model are summarized in Table 5.2, and depicted in Figure 5.2C and D. To test the difference in the group-level distribution of the inverse temperature parameter, we calculated the 95% HDI of the difference in mean parameters $\mu_{\beta,\text{Other}}$ - $\mu_{\beta,\text{Self}}$ Figure 5.2D. The resulting HDI was entirely contained within the positive numbers, indicating generally higher inverse temperature parameters during prosocial learning.

In summary, we found that participants' choices where characterized by higher inverse temperature parameters during prosocial compared to self-relevant learning. However, we found no differences in learning rates between the two conditions. This suggests that participants were more sensitive to differences in subjective values when making a choice for the other person than when making a choice for themselves. At the same time, the degree to which participants updated their subjective values in response to outcomes did not differ between self-relevant and prosocial learning contexts.

5.3.3 Associations between value sensitivity and optimal choices

As computational modeling had revealed that participants were more sensitive to value differences during prosocial choices than self-relevant choices, we next tested if this finding could explain the higher number of optimal choices for the other than for oneself. As a simple index of this association, we correlated the difference in number of optimal choices (*Other - Self*) with the difference in value sensitivity (β_{Other} - β_{Self}). This revealed a significant correlation (r = 0.476, $p < 0.001$). To investigate the role of value sensitivity in more detail, we conducted multilevel mediation analysis. This revealed that the effect of *condition* (*Other vs. Self*) on the number of optimal choices was significantly mediated by value sensitivity (indirect effect = 0.059, 95% BCa-CI = [0.044, 0.077]). Since the direct effect of *condition* ceased to be significant (direct effect = 0.014, p = 0.666), this indicated full mediation. In summary, the better performance during prosocial learning (compared to self-relevant learning) could be explained

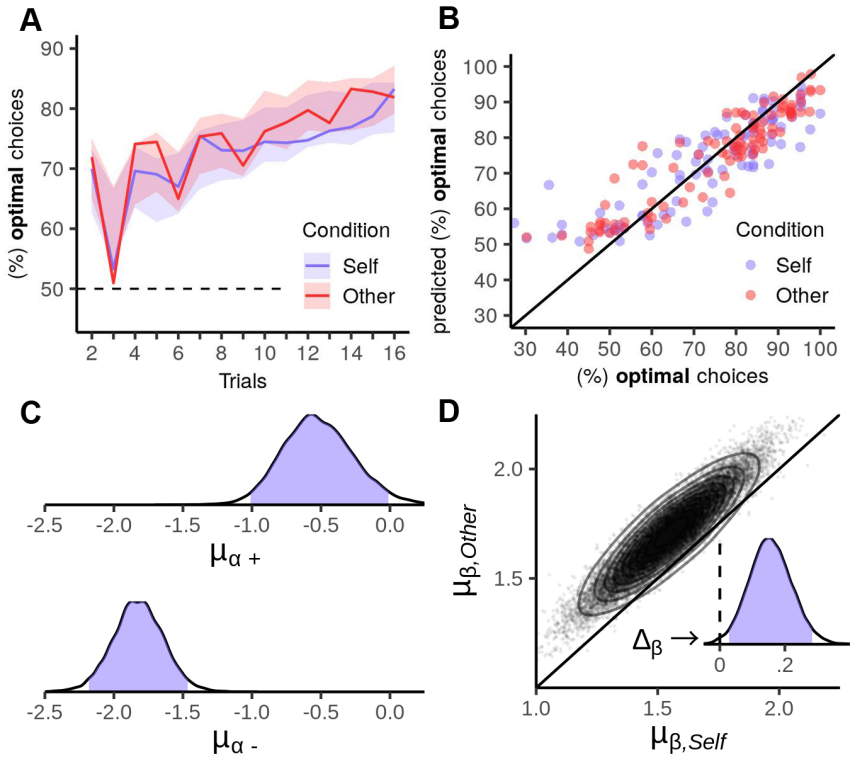

FIGURE 5.2: Prosocial learning task: behavioral results and model parameters. **A)** Percentage of optimal choices per trial. Solid lines depict the percentage of times participants chose the symbol that minimized the expected number of painful stimuli. Overall, participants made significantly more optimal choices during prosocial learning (Other, red) than during self-relevant learning (Self, blue). Shaded areas depict 95% highest-density intervals from the posterior predictive distribution of optimal choices derived from the computational model. The posterior predictive distribution matches the actual responses well, indicating that the model was able to capture participants' learning curves. **B)** Posterior predictive check for percentage of optimal choices. The real percentage of optimal choices of each participant is plotted against the average percentage of optimal choices predicted by the winning model. The high congruence between actual and predicted responses suggests that good absolute model fit was achieved, but note that the model appears to overestimate the percentage of optimal choices for those participants who performed below the chance level of 50%. **C)** Posterior distributions of group level means of the learning rates for positive ($\mu_{\alpha+}$) and negative ($\mu_{\alpha-}$) outcomes. The curves depict the probability density describing the posterior distribution of the mean parameters. The blue-shaded area depicts the 95% highest-density-interval (the interval containing the 95% of the values with the highest posterior probability). Note that the mean parameters are depicted on the logit scale. **D)** Joint posterior distribution of the group level means of the inverse temperature parameters, $\mu_{\beta,\text{Self}}$ and $\mu_{\beta,\text{Other}}$. Points depict Monte Carlo Markov Chain (MCMC) samples, contours depict the estimated joint density. The inserted plot depicts the posterior distribution of the difference $\mu_{\beta,\text{Other}} - \mu_{\beta,\text{Self}}$. The blue-shaded area depicts the 95% highest-density-interval. The dashed vertical line marks zero.

TABLE 5.2: Parameter estimates of the computational model. The best-fitting computational model (M4.2) allowed for different learning rates for positive outcomes (α+) and negative outcomes (α-), and for different inverse temperature parameters for self-oriented learning (β_{Self}) and prosocial learning (β_{Other}). Posterior mean: mean of the posterior distribution of the respective parameter. 95% HDI: Highest-density interval. Posterior mean (transformed): mean of the posterior distribution after transformation (logistic function for learning rates; exponential function for inverse temperature).

Parameter	Posterior mean	95% HDI	Posterior mean (transformed)	95% HDI (transformed)
$\mu_{\alpha+}$	-0.53	[-1.04 -0.01]	0.37	[0.26, 0.49]
$\sigma_{\alpha+}$	0.92	[0.12, 1.64]		
$\mu_{\alpha-}$	-1.83	[-2.19, -1.46]	0.14	[0.10, 0.18]
$\sigma_{\alpha-}$	0.92	[0.62, 1.25]		
$\mu_{\beta,\text{Self}}$	1.54	[1.20, 1.86]	4.73	[3.27, 6.33]
$\sigma_{\beta,\text{Self}}$	1.12	[0.45, 1.76]		
$\mu_{\beta,\text{Other}}$	1.70	[1.36, 2.02]	5.53	[3.82, 7.39]
$\sigma_{\beta,\text{Other}}$	1.12	[0.45, 1.76]		
$\mu_{\beta,\text{Other}}$ - $\mu_{\beta,\text{Self}}$	0.16	[0.03, 0.29]	0.80	[0.11, 1.54]

by participants being more sensitive to the subjective values of the symbols during prosocial decisions.

5.3.4 Association between prosocial choice behavior and trait empathy

Since we found higher value-sensitivity during prosocial compared to self-relevant learning, we next investigated whether this prosocial preference was associated with empathic traits. We regressed the difference in individual inverse temperature parameter estimates (β_{Other} - β_{Self}) on the five subscales of the QCAE. All variance-inflation factors were below 2, indicating negligible multicollinearity between predictors. We found that the subscales explained a significant amount of variance of this difference score (R^2 = 0.183, $F_{(5,81)}$ = 3.62, p = 0.005). On the level of single predictors, *emotional contagion* had a significant positive association with the difference score (standardized regression coefficient = 0.43, SE = 0.12, $t_{(81)}$ = 3.614, $p < 0.001$). This subscale measures how strongly individuals automatically share the emotions of others. Moreover, *proximal responsivity* had a significant negative association with the difference score (standardized regression coefficient = -0.36, SE = 0.13, $t_{(81)}$ = -2.735, p = 0.008). This subscale consists of items that measure the tendency to become upset in response to other individuals' problems. No other subscale was a significant predictor (all $p > 0.05$, see Table 5.3). Thus, participants scoring higher in emotional contagion showed higher value sensitivity during prosocial compared to self-relevant learning, and participants who reported a higher tendency towards proximal responsivity displayed a smaller difference in value sensitivity during prosocial vs. self-relevant learning.

5.3.5 Brain activation related to valuation

As we found that participants were more sensitive to differences in subjective values during prosocial learning compared to self-relevant learning, we next investigated whether differences in neural valuation processes might underpin this finding. As a

TABLE 5.3: Regression of difference in value sensitivity (prosocial learning vs. self-oriented learning) on empathic traits. Dependent variable: Difference between inverse temperature parameters ($\beta_{\text{Other}} - \beta_{\text{Self}}$). VIF = variance inflation factor.

Predictor	**Standardized Regression weight**	**VIF**	$t_{df=81}$	**p-Value**
Intercept	-0.066 (0.099)	-	-0.666	0.507
Perspective Taking	0.133 (0.110)	1.16	1.210	0.230
Online Simulation	-0.090 (0.114)	1.30	-0.784	0.436
Emotional Contagion	0.433 (0.120)	1.40	3.614	<0.001
Proximal Responsivity	-0.360 (0.132)	1.65	-2.735	0.008
Peripheral Responsivity	-0.097 (0.114)	1.25	-0.850	0.398

manipulation check, we tested if the brain regions typically associated with valuation processes where also activated during our task, regardless of condition (*Self* vs. *Other*). For this, we assessed which areas exhibited activation patterns that where correlated with the trial-wise value difference between the chosen symbol and the unchosen symbol (Δ_{Value}), as derived from the computational model. This revealed a significant positive association between value difference and brain activity in anterior and posterior midline structures, including the VMPFC, subgenual anterior cingulate cortex (sgACC), and precuneus, as well as the right hippocampus and the bilateral middle temporal gyrus (contrast $\Delta_{\text{Value}} > 0$; Figure 5.3A, Table 5.4). Next, we tested if valuation-related brain activity was increased during prosocial compared to self-relevant choices, or vice versa. However, we found no differences in average activation between prosocial and self-relevant valuation processes (contrast $\Delta_{\text{Value, Other}} > \Delta_{\text{Value, Self}}$; contrast $\Delta_{\text{Value, Self}} > \Delta_{\text{Value, Other}}$).

We next tested if the individual differences in value sensitivity (identified by the modeling of the behavioral data) could be explained by differences in brain activity underpinning valuation. For this, we correlated the individual estimates of the contrast [$\Delta_{\text{Value, Other}} > \Delta_{\text{Value, Self}}$] with the difference in inverse temperature parameters (β_{Other} - β_{Self}). This revealed clusters of positive correlations within the VMPFC , the precuneus and the left angular gyrus (Figure 5.3B and C, Table 5.4).

In summary, we found that the difference in the subjective values of possible actions was positively associated with activity within the VMPFC and sgACC, the precuneus, and the middle temporal gyrus. Moreover, participants who made more value-sensitive choices for the other compared to for themselves also exhibited increased valuation-related activity within these brain areas during prosocial choices.

5.3.6 Functional connectivity during choices

Since the VMPFC is an area that has consistently been linked to valuation processes in previous research, we focused on connectivity of this area for answering the question whether valuation related modulation also mediated the communication between brain regions. We found significantly increased functional connectivity of the VMPFC with the posterior right middle temporal gyrus/angular gyrus when choosing for the other than when choosing for oneself (contrast Other > Self; Figure 5.3D; Table 5.4. The cluster corresponded to an area that has been referred to as right temporoparietal junction (rTPJ; Quesque & Brass, 2019; Schurz et al., 2017; Silani et al., 2013). The reverse contrast (Self > Other) revealed no significant connectivity differences.

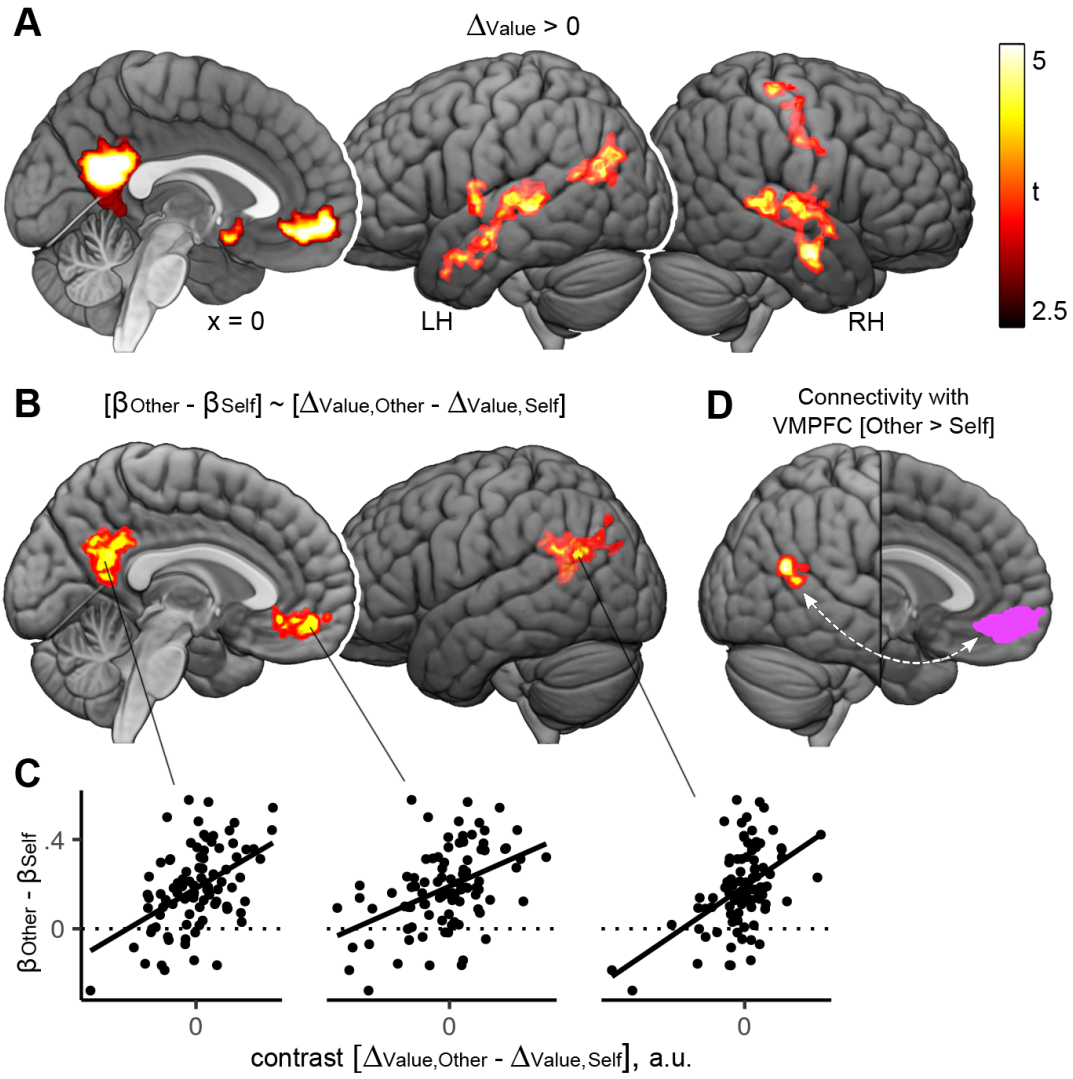

FIGURE 5.3: Results of fMRI analyses of brain activity and connectivity during choices in the prosocial learning task. **A)** Brain activity associated with valuation processes during choices. Depicted are brain areas where activity was significantly and positively correlated with the difference in subjective values of the symbols (Δ_{Value}: value of chosen symbol minus value of unchosen symbol). In these areas, activity was high when the difference between subjective values was high; and low, when the value difference was small. **B)** Brain activity correlated with interindividual differences in prosocial value sensitivity. Depicted are brain areas where the contrast $[\Delta_{\text{Value,Other}} - \Delta_{\text{Value,Self}}]$ was significantly correlated with the difference in value sensitivity for other vs. self ($\beta_{\text{Other}} - \beta_{\text{Self}}$). Participants with higher valuation-related activity in these areas during prosocial learning (compared to self-oriented learning) were also more sensitive to values relevant for the other compared to values relevant for self (as indicated by computational modeling of the behavioral data). **C)** Scatter plots illustrating the correlation between $[\beta_{\text{Other}} - \beta_{\text{Self}}]$ and $[\Delta_{\text{Value,Other}} - \Delta_{\text{Value,Self}}]$ at the clusters depicted in B. Values were extracted from peak voxels. **D)** Results of the generalized psychophysiological interaction (gPPI) analysis. A cluster in the right temporoparietal junction shows increased connectivity with the ventromedial prefrontal cortex (VMPFC, displayed in pink) during choices for the other, compared to choices for self. Note: all results $p < 0.05$ FWE-corrected on the cluster-level, cluster-defining threshold $p < 0.001$ (see also Table 5.4). L: left, R: right, H: hemisphere, VMPFC: ventromedial prefrontal cortex.

TABLE 5.4: Brain activity and functional connectivity during valuation. We report the first local maximum within each cluster. Effects were tested for significance using cluster inference with a cluster defining threshold of $p < 0.001$ and a cluster probability of $p < 0.05$, FWE-corrected. gPPI = generalized psychophysiological interaction analysis. VMPFC = ventromedial prefrontal cortex. *For this effect, we list the 10 clusters with the highest peak z-values.

	MNI Coordinates				
Contrast and Brain Region	**x**	**y**	**z**	**z-value**	**Cluster size**
Parametric modulation by Δ value*					
L/R Precuneus	4	-52	18	6.57	1869
L/R superior frontal gyrus, medial orbital	0	58	-6	5.85	991
L middle temporal gyrus	-62	-32	2	5.65	1635
R superior temporal gyrus	60	-26	2	5.42	1363
R fusiform gyrus	36	-52	-4	5.22	115
L hippocampus	-20	-16	-20	5.06	61
L angular gyrus	-46	-70	24	4.97	505
L/R olfactory cortex	0	8	-14	4.48	78
L superior temporal gyrus	-34	-22	2	4.33	60
R postcentral gyrus	48	-24	62	4.30	72
Correlation $[\Delta_{\text{Value,Other}} - \Delta_{\text{Value,Self}}] \times [\beta_{\text{Other}} - \beta_{\text{Self}}]$					
L/R Precuneus	6	-54	24	4.82	655
L/R superior frontal gyrus, medial orbital	2	52	-10	4.60	341
L angular gyrus	-40	-64	24	4.49	206
gPPI: Functional connectivity with VMPFC (Other >Self)					
R middle temporal gyrus	46	-64	20	4.47	93

5.3.7 Brain activation related to outcomes and prediction errors

So far, we found that differences between prosocial and self-relevant learning were visible in the domain of valuation (more optimal choice behavior and higher value sensitivity during prosocial compared to self-relevant learning), but not in the domain of outcome evaluation (same learning rates for prosocial and self-relevant learning). We therefore did not predict that fMRI analyses of outcome events to be informative for our research question. Here, we briefly report the basic results of these analyses, and refer the reader to Supplementary Table 2.A.1 and Figure 5.4 for detailed results.

Analyzing brain activity associated with the presentation of outcomes, we found that positive outcomes engaged areas typically associated with reward processing, such as the ventral striatum and VMPFC, to a stronger degree than negative outcomes (contrast Positive > Negative, Figure 5.4A, Table Supplementary Table 2.A.1). These effects were similar for self-relevant and prosocial learning. Negative outcomes, in contrast, led to increased activity in bilateral AI, AMCC and supramarginal gyrus (contrast Negative > Positive, Figure 5.4A, Supplementary Table 2.A.1), and did so to a greater extent during self-relevant than during prosocial learning (contrast (Negative – Positive)$_{\text{Self}}$ > (Negative – Positive)$_{\text{Other}}$, Figure 5.4B, Supplementary Table 2.A.1). We further analyzed correlations between brain activity and absolute prediction errors derived from the computational model, indicating processes that were specific to unexpected outcomes. Brain areas that were specifically activated by positive and negative prediction errors are listed in Supplementary Table 2.A.2. Notably, we found that prediction errors engaged areas within the bilateral dorsomedial prefrontal cortex and the right orbitofrontal cortex to a greater degree during prosocial learning than during self-relevant learning. We found no brain areas where activity associated with prediction errors was greater during self-relevant than prosocial learning.

5.4 Discussion

Are humans as good at learning to avoid harm to others as they are at learning to avoid self-harm? And if so, which neural processes underpin these two types of learning? Here, we find that participants were, in fact, *better* at protecting another person from pain than themselves. The higher number of optimal choices for the other was explained by participants being more sensitive to the values of choice options during prosocial, compared to self-relevant learning, and this prosocial preference was significantly associated with empathic traits. On the neural level, higher sensitivity for other-related values was reflected by a stronger engagement of the VMPFC during valuation. Moreover, the VMPFC exhibited a higher connectivity to the right TPJ during choices affecting the other. Taken together, these findings suggest that humans are particularly adept at learning to protect others. This ability appears implemented by neural mechanisms that overlap with those supporting self-relevant learning, but with the additional recruitment of structures relevant in social cognition.

Previous evidence suggested that humans are inherently egocentric learners, adapting their behavior faster for their own benefit than for others' (Kwak et al., 2014; Lockwood et al., 2016). Importantly, these studies investigated prosocial learning in experimental settings which used monetary outcomes. Our data indicate that the situation might be reversed in the context of harm-avoidance. A similar "hyperaltruistic" tendency was reported by Crockett et al. (2014, 2015), who found that individuals chose to spend more money to protect another person from pain than themselves. Our

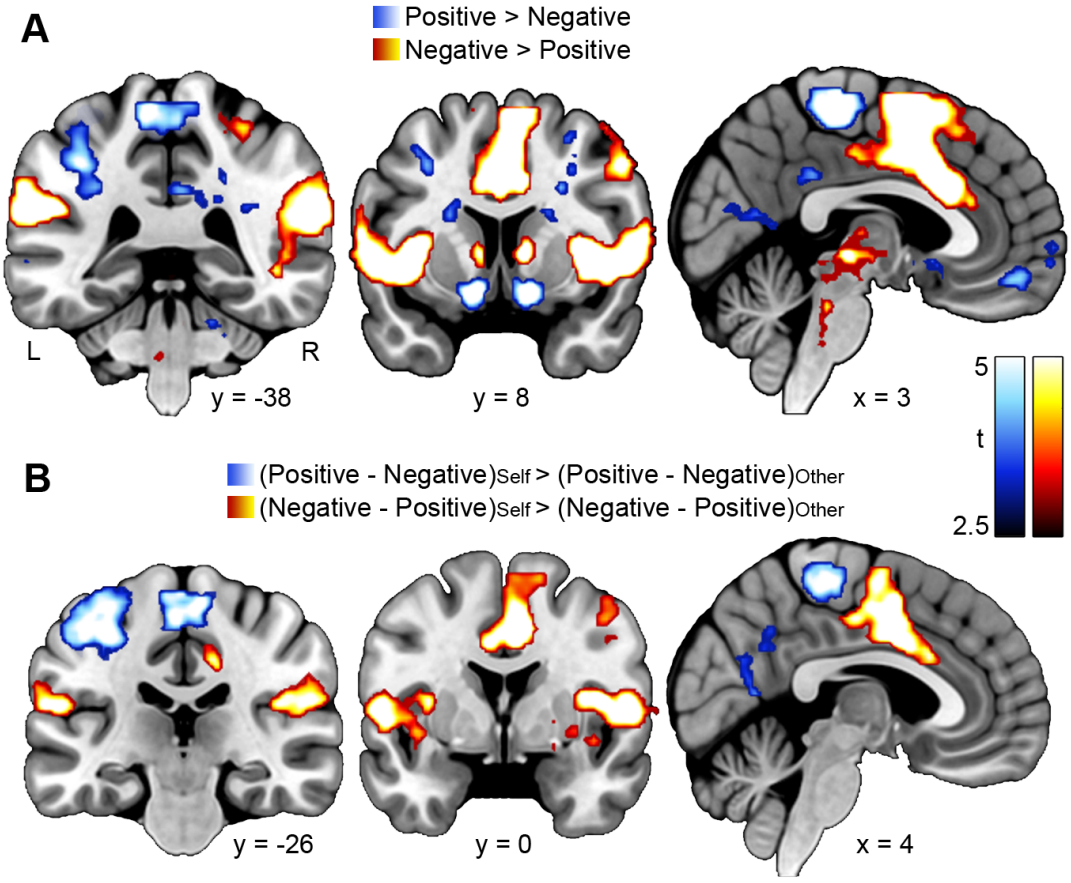

FIGURE 5.4: Results of fMRI analyses of brain activity during the presentation of outcomes in the prosocial learning task. A) Effects of outcome valence. Blue-to-white clusters delineate areas with increased activity after a positive outcome, i.e. trials in which the choice made would not be followed by painful stimulation. Red-to-yellow clusters indicate areas with increased activity after a negative outcome, i.e. trials in which the choice made would be followed by painful stimulation. B) Interaction between outcome valence (positive vs. negative) and condition (Self vs. Other). Blue-to-white clusters indicate areas where the effect of positive outcomes was stronger during self-relevant compared to prosocial learning. Red-to-yellow clusters indicate areas where the effect of negative outcomes was stronger during self-relevant compared to prosocial learning. Note: all results $p < 0.05$ FWE-corrected based on a cluster-defining threshold of $p < 0.001$. Results are listed in Supplementary Table 2.A.1. Additional results regarding the parametric modulation of outcome-related brain activity by prediction errors are listed in Supplementary Table 2.A.2. L: left, R: right.

study provides a crucial extension to these findings, demonstrating that individuals act "hyperaltruistically" not only in situations where they can explicitly weigh self- and other-relevant outcomes, but also when the consequences of their behavior have to be learned implicitly. This lends essential support to the notion that prosociality may be an intuitive act (Zaki & Mitchell, 2013).

The higher number of optimal choices for the other was well explained by a greater sensitivity to subjective values during prosocial compared to self-relevant learning. The rate with which participants updated subjective values was not different between the two conditions, though. This implies that prosocial learning differs from self-relevant learning not in the way that information is collected, but in how strongly this information is weighed when choosing between actions. Moreover, we found that this prosocial "bonus" in value-sensitivity was related to empathic traits. Participants who reported to strongly share the emotions of others (subscale *Emotional Contagion* of the QCAE; Reniers et al., 2011) were also more sensitive to other-related values, implying that individuals who more readily feel what others feel are also more sensitive to the possible consequences of their actions for others. However, the tendency to get distressed by other people's problems (*Proximal Responsitivity*) appeared detrimental to sensitive valuation, perhaps due to the negative effect of stress on executive functions (Shields et al., 2016). This aligns with the notion that empathic sharing of emotions, but not personal distress, is a key driver of prosocial behavior (Decety et al., 2016; Hein et al., 2010).

In line with previous research (Kable & Glimcher, 2009; Ruff & Fehr, 2014), we found that the VMPFC was engaged during both prosocial and self-relevant valuation. Interestingly, we observed that differences in value-sensitivity on the behavioral level were correlated with differences in valuation-related brain activity: Participants who were more sensitive to values during prosocial learning (compared to self-relevant learning) also displayed increased valuation-related VMPFC activity during prosocial choices. Differences in prosocial learning performance may thus be partially explained by differences in VMPFC involvement during valuation, possibly because strongly activated value representations facilitate the readout of values (Grueschow et al., 2015). Another important finding was that the VMPFC exhibited a stronger coupling with the right TPJ (rTPJ) during prosocial choices. The rTPJ has been extensively linked to self-other distinction and related social functions (Quesque & Brass, 2019; Schurz et al., 2017; Silani et al., 2013). One function of this area is perspective-taking (Lamm et al., 2016; Schurz et al., 2017; Silani et al., 2013), which also has direct effects on prosocial decision-making (Morishima et al., 2012). Thus, prosocial valuation in our task may be informed by processes that enable participants to take the confederate's perspective, for example by a mapping of the choice options to their consequences for the other (Hare et al., 2010; Janowski et al., 2013). A complementary interpretation is that the rTPJ engagement reflects the higher demand for self-other distinction during prosocial learning (Lamm et al., 2016; Quesque & Brass, 2019). For instance, the valuation processes in the VMPFC may call for a stronger signalling that the current decision will affect another person, and not oneself (Carter & Huettel, 2013; Tomova et al., 2020).

While these findings suggest a central role of the VMPFC in successfully learning to avoid others' harm, they cannot fully explain the generally higher number of optimal choices during prosocial learning. We did not find higher average activity during prosocial valuation, precluding claims that our sample's better performance for the other could be simply explained by generally stronger VMPFC engagement. It has

been shown, though, that the VMPFC can switch dynamically between different frames of reference, such as values relevant for self vs. other (Nicolle et al., 2012). This mechanism may mask differences in average activity, making them too subtle to be identified with GLM-based analyses. Further studies tailored for more complex analysis approaches, such as multivariate pattern analysis (Norman et al., 2006) or joint brain-behavior modeling (Palestro et al., 2018) might further elucidate these issues.

Based on the mathematical role of the inverse temperature parameter (see **Methods**), we have interpreted β as an indicator of value-sensitivity. Importantly, our findings are also consistent with the notion that participants showed less exploratory tendencies during prosocial choices (Cinotti et al., 2019; Humphries et al., 2012), perhaps due to a higher risk-aversion when choices might harm the other (Batteux et al., 2019). Future studies could thus optimize their design to differentiate between the roles of valuation and exploration/exploitation during prosocial decisions.

With regards to outcome evaluation, participants weighted positive prediction errors more strongly than negative prediction errors, implying that they tended to learn through successful pain-avoidance, rather than through pain itself (Eldar et al., 2016). This result is consistent with recent findings suggesting that higher learning rates for positive outcomes reflect a confirmation bias in learning (Lefebvre et al., 2017; Palminteri, Lefebvre, et al., 2017) (but see Gershman, 2015, for contrasting results). Surprisingly, we found no behavioral evidence that prosocial and self-relevant learning differed in terms of outcome evaluation, which contrasts previous studies of reward or associative learning that find lower learning rates for others (Lockwood et al., 2016, 2018). We did, however, observe significant differences in brain activity evoked by outcomes: while negative outcomes engaged bilateral AI and ACC reliably during prosocial and self-relevant learning, responses where stronger for self-relevant outcomes. This might indicate stronger anticipation of pain (Ploghaus et al., 1999), as self-related negative outcomes also signalled the imminent delivery of painful stimulation. As we found no corresponding difference in behavior, it is possible that the stronger aversive responses did not reflect processes involved in the updating of values.

Several limitations of our study need to be addressed. We only tested men, which possibly limits our conclusions to males. Six out of 96 participants reported doubts about the confederate after the second session of the overarching project (i.e., two weeks after the session analysed here). We have no information on whether these doubts were already present during the first session, which precludes to assess if the observed effects were different in participants who doubted the cover story. It seems unlikely though that such a small part of the sample fundamentally affects the results.

Furthermore, participants were instructed that the confederate would know about the outcomes of their choices during both conditions of the task. Thus, it is possible that egocentric motivations, such as care for reputation, contributed to the difference between prosocial and self-relevant learning, as harming others might be seen as more detrimental to social status than harming oneself. We believe that the substantial correlations between prosocial learning performance and empathic traits point towards an important role of prosociality in learning to avoid other's harm. To clearly disentangle the (potentially complementary) roles of "selfish" and prosocial motives, further studies in which social observation is experimentally manipulated are needed. Furthermore, our findings cannot fully answer the question whether the prosocial

tendencies we observed are specific to situations in which the other's physical integrity is at risk, or if similar levels of prosociality would be displayed in situations where the other's "harm" is defined through financial loss. Indeed, previous studies directly comparing selfish tendencies in the domains of monetary rewards compared to protecting others from losses have found that people have a similar selfish bias in both contexts (Lockwood et al., 2017). To examine this further, future studies should vary the modality of harm (e.g.., pain, financial loss) within the same design. Finally, our task reflected a situation in which optimal behavior for the other did not preclude optimal behavior for oneself. In many real-life situations, minimizing others' harm and minimizing self-harm are conflicting objectives, with examples ranging from in-vivo organ-donation (Marsh et al., 2014) to war. Further investigating how such conflicts influence prosocial learning would be of major importance for future research.

In conclusion, our findings support the notion that humans are not always "selfish" learners. Rather, they appear to be particularly good at learning for the benefit of other people when their physical integrity, rather than their financial success, is at risk. Our findings have important implications for future research on prosocial behavior: On the one hand, they highlight the importance of considering different outcomes when investigating learning for others; on the other, they demonstrate that we may show "intuitive prosociality" (Zaki & Mitchell, 2013) not only during explicit decision-making (Crockett et al., 2014, 2015), but also when prosocial actions about harm avoidance depend on implicit learning.

Acknowledgments

This work was funded by the Vienna Science and Technology Fund (WWTF VRG13-007). We would like to thank Lei Zhang and Nace Mikus for helpful discussions concerning the modeling analyses, and feedback on the manuscript, and David Sastre-Yagüe, Gloria Mittmann, André Lüttig, Sophia Shea and Leonie Brög for assistance during data collection.

Conflict of Interest

The authors declare no competing financial interests.

5.5 Appendix

5.A Detailed Description of the Hierarchical Reinforcement Learning Models

Our approach takes inspiration from the literature on generalized linear mixed effects models (GLMM; f.e. McCulloch & Neuhaus, 2005) and software for the Bayesian estimation of such models, such as the R package *brms* (Bürkner, 2017). In particular, we adopt the following terminology of GLMM:

- We define as *fixed effect* that part of the effect of a factor which applies to the whole population (i.e., the expected value of the effect).
- We define as *random effect* that part of an effect that is specific to a certain observatory unit, in our case a participant.

Also see Daw (2011) for the use of this terminology in hierarchical reinforcement learning models. As in GLMM, we will assume that any participant-level parameter θ_i follows a normal distribution, $\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta)$, characterized by some population-level mean μ_θ and standard deviation σ_θ. In cases where θ_i describes the effect of an experimental factor, we can equate μ_θ with the fixed effect, and σ_θ with the standard deviation of the associated random effect. However, we must emphasize that, in contrast to GLMMs, the models we describe here are decidedly non-linear (in the sense that there is no transformation which achieves that the expected value of the transformed data is a linear function of the parameters). In the following, we will first describe the parameterization of our model, and then describe and justify the priors we used.

Parameterization

Participant-level model

We consider an extension to the simple reinforcement learning (RL) model (Sutton & Barto, 2018). We will first give a short description of the simple RL model, and then focus on its extension. Let i index participants and t index trials. Further, let C_{it} be a dichotomous variable indicating the choice of participant i in trial t, with $C_{it} = 0$ if symbol A was chosen, and $C_{it} = 1$ if symbol B was chosen. Also, let R_{it} be a dichotomous variable indicating the outcome delivered to participant i in trial t, with $R_{it} = 0$ if a negative outcome was delivered, and $R_{it} = 1$ if a positive outcome was delivered. Finally, let $\vec{V}_{it}$ be a vector of length 2, containing the subjective values of the two symbols for participant i in trial t, with the value of A as the first component, and the value of B as the second component. According to the simple RL model, the choice of participant i in trial t is then distributed as

$$P(C_{it}|\beta_i, \vec{V}_{it}) = \frac{1}{1 + \exp(-\beta_i \Delta(\vec{V}_{it}, C_{it}))}, \tag{RL1}$$

where β_i is the inverse temperature parameter of participant i, and $\Delta(\vec{V}_{it}, C_{it})$ is the difference in values between the chosen and the unchosen symbol, calculated as the dot product

$$\Delta(\vec{V}_{it}, C_{it}) = \begin{pmatrix} 1 - 2C_{it} \\ 2C_{it} - 1 \end{pmatrix} \cdot \vec{V}_{it}.$$

The inverse temperature β_i is restricted to be a positive real number, reflecting the assumption that participants will not systematically choose the symbol with the lower subjective value. Further, the simple RL model assumes that the subjective value of the chosen symbol is updated by the weighted prediction error, while the subjective value of the unchosen symbol remains unchanged. This can be written concisely as

$$\vec{V}_{i,t+1} = \vec{V}_{it} + \alpha_i \begin{pmatrix} 1 - C_{it} & 0 \\ 0 & C_{it} \end{pmatrix} \left(\begin{pmatrix} R_{it} \\ R_{it} \end{pmatrix} - \vec{V}_{it} \right), \tag{RL2}$$

where α_i is the learning rate parameter of participant i, and restricted to the interval $(0, 1)$. Initializing $\vec{V}_{i0}$ to some value, for example $(0.5, 0.5)'$, completes the simple RL model.

We now consider two extensions of the model. First, we will consider a model in which different learning rate parameters α^+ and α^- govern the value update after positive and negative outcomes, respectively. This can be achieved by slightly changing formula **RL2** to

$$\vec{V}_{i,t+1} = \vec{V}_{it} + (R_{it}\alpha_i^+ + (1 - R_{it})\alpha_i^-) \begin{pmatrix} 1 - C_{it} & 0 \\ 0 & C_{it} \end{pmatrix} \left(\begin{pmatrix} R_{it} \\ R_{it} \end{pmatrix} - \vec{V}_{it} \right).$$

Second, we will consider models in which some of the participant-level parameters may differ between the two levels of an experimental factor. For this, we introduce another index j for the level of the experimental factor. (In the case of our article, the experimental factor is learning for *Self* vs. learning for *Other*, and may be coded as $j = 1$ for *Self* and $j = 2$ for *Other*). Models allowing for an effect of the experimental factor in a certain parameter are then simply specified by adding this index to the relevant quantities. Thus, our extended reinforcement learning model (extRL) is specified by

$$P(C_{ijt}|\beta_{ij}, \vec{V}_{ijt}) = \frac{1}{1 + \exp(-\beta_{ij}\Delta(\vec{V}_{ijt}, C_{ijt}))} \tag{extRL1}$$

and

$$\vec{V}_{ij,t+1} = \vec{V}_{ijt} + (R_{ijt}\alpha_{ij}^+ + (1 - R_{ijt})\alpha_{ij}^-) \begin{pmatrix} 1 - C_{ijt} & 0 \\ 0 & C_{ijt} \end{pmatrix} \left(\begin{pmatrix} R_{ijt} \\ R_{ijt} \end{pmatrix} - \vec{V}_{ijt} \right), \tag{extRL2}$$

where the added index j indicates which level of the experimental factor the respective quantity is associated with. For example, β_{ij} is the inverse temperature parameter of participant i in level j of the experimental factor. Note that **extRL1** and **extRL2** define the *full* model allowing for experimental effects in all free parameters. All models we tested are nested within this full model, and can be derived by putting restrictions on the parameter space. For example, the simple RL model described above is obtained from the extended model by adding the restrictions that $\alpha_{i1}^+ = \alpha_{i2}^+ = \alpha_{i1}^- = \alpha_{i2}^-$ and $\beta_{i1} = \beta_{i2}$ for all i.

Reparameterization to unrestricted space

The model formulation described above necessitates that the free parameters are restricted to certain subsets of the real numbers, in particular $\alpha_{ij}^+, \alpha_{ij}^- \in (0, 1)$ and $\beta_{ij} \in (0, \infty)$. For hierarchical modeling, though, it is more convenient to reparameterize

the model in such a way that the parameters are defined on the whole field of real numbers. For the inverse temperature, this can be achieved by defining the parameter $\tilde{\beta}_{ij}$ as the logarithm of β_{ij} in **extRL1**. For the learning rates, we may define $\tilde{\alpha}_{ij}^{+}$ and $\tilde{\alpha}_{ij}^{-}$ as the logits of α_{ij}^{+} and α_{ij}^{-} in **extRL2**. The reparameterization is thus achieved by simply defining the quantities in **extRL1** and **extRL2** as

$$\begin{aligned} \beta_{ij} &= \exp(\tilde{\beta}_{ij}), \\ \alpha_{ij}^{+} &= \frac{1}{1+\exp(-\tilde{\alpha}_{ij}^{+})}, \\ \text{and } \alpha_{ij}^{-} &= \frac{1}{1+\exp(-\tilde{\alpha}_{ij}^{-})}. \end{aligned}$$

Hierarchical model

Per our model, the behavior of a single participant i is governed by the vector of parameters $\vec{\theta}_i = (\tilde{\alpha}_{i1}^{+}, \tilde{\alpha}_{i1}^{-}, \tilde{\beta}_{i1}, \tilde{\alpha}_{i2}^{+}, \tilde{\alpha}_{i2}^{-}, \tilde{\beta}_{i2})'$ (note that the components of this vector might be subject to equality restrictions when we consider a restricted model). We now seek to model how $\vec{\theta}_i$ is distributed in the population. As in GLMM, we will assume that $\vec{\theta}_i$ follows a multivariate normal distribution, $\vec{\theta}_i \sim \mathcal{N}(\vec{\mu}_\theta, \Sigma_\theta)$ for some vector $\vec{\mu}_\theta$ and covariance matrix Σ_θ. As we are primarily interested in the effects of the experimental factor, it is convenient to reparameterize $\vec{\theta}_i$ in terms of the *means* of parameters across the two levels of the factor, which we will denote by $m_i^{(.)}$ (with the superscript indicating the respective parameter); and the *difference* between the two levels, which we will denote by $d_i^{(.)}$. This reparameterization is achieved by the linear transformation

$$\begin{pmatrix} m_i^{\alpha+} \\ m_i^{\alpha-} \\ m_i^{\beta} \\ d_i^{\alpha+} \\ d_i^{\alpha-} \\ d_i^{\beta} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_{i1}^{+} \\ \tilde{\alpha}_{i1}^{-} \\ \tilde{\beta}_{i1} \\ \tilde{\alpha}_{i2}^{+} \\ \tilde{\alpha}_{i2}^{-} \\ \tilde{\beta}_{i2} \end{pmatrix},$$

and the original parameters can be obtained by the inverse transformation

$$\begin{pmatrix} \tilde{\alpha}_{i1}^{+} \\ \tilde{\alpha}_{i1}^{-} \\ \tilde{\beta}_{i1} \\ \tilde{\alpha}_{i2}^{+} \\ \tilde{\alpha}_{i2}^{-} \\ \tilde{\beta}_{i2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & -\frac{1}{2} \\ 1 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} m_i^{\alpha+} \\ m_i^{\alpha-} \\ m_i^{\beta} \\ d_i^{\alpha+} \\ d_i^{\alpha-} \\ d_i^{\beta} \end{pmatrix}.$$

As a linear function of a normally distributed random vector, the new vector of parameters is normally distributed as well. To reduce complexity, we will further assume that the subvector $\vec{m}_i = (m_i^{\alpha+}, m_i^{\alpha-}, m_i^{\beta})'$ is independent of the subvector $\vec{d}_i = (d_i^{\alpha+}, d_i^{\alpha-}, d_i^{\beta})'$. In other words, we assume that the effects of the experimental factor on the parameters are not correlated with the mean values of the parameters.

Under this assumption, we may write the distribution of $\vec{m}_i$ as

$$\begin{pmatrix} m_i^{\alpha+} \\ m_i^{\alpha-} \\ m_i^{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_m^{\alpha+} \\ \mu_m^{\alpha-} \\ \mu_m^{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_m^{\alpha+} & 0 & 0 \\ 0 & \sigma_m^{\alpha-} & 0 \\ 0 & 0 & \sigma_m^{\beta} \end{pmatrix} R_m \begin{pmatrix} \sigma_m^{\alpha+} & 0 & 0 \\ 0 & \sigma_m^{\alpha-} & 0 \\ 0 & 0 & \sigma_m^{\beta} \end{pmatrix} \right), \text{(HM1)}$$

where $\mu_m^{\alpha+}$, $\mu_m^{\alpha-}$ and μ_m^{β} are the population-level means of the parameters, $\sigma_m^{\alpha+}$, $\sigma_m^{\alpha-}$ and σ_m^{β} are the population-level standard deviations, and R_m is a correlation matrix of size 3. Similarly, we can write the distribution of $\vec{d}_i$ as

$$\begin{pmatrix} d_i^{\alpha+} \\ d_i^{\alpha-} \\ d_i^{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_d^{\alpha+} \\ \mu_d^{\alpha-} \\ \mu_d^{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_d^{\alpha+} & 0 & 0 \\ 0 & \sigma_d^{\alpha-} & 0 \\ 0 & 0 & \sigma_d^{\beta} \end{pmatrix} R_d \begin{pmatrix} \sigma_d^{\alpha+} & 0 & 0 \\ 0 & \sigma_d^{\alpha-} & 0 \\ 0 & 0 & \sigma_d^{\beta} \end{pmatrix} \right)$$

where $\mu_d^{\alpha+}$, $\mu_d^{\alpha-}$ and μ_d^{β} are the population-level means of the effects of the experimental factor on the parameters, $\sigma_d^{\alpha+}$, $\sigma_d^{\alpha-}$ and σ_d^{β} are the population-level standard deviations of the effects, and R_d is a correlation matrix of size 3. However, to facilitate the formulation of appropriately informed priors (see below), we choose to reparameterize the distribution of $\vec{d}_i$ in the following way:

- Let $\delta^{(\cdot)} := \frac{\mu_d^{(\cdot)}}{\sigma_d^{(\cdot)}}$ be the *effect size* of the experimental factor with regards to the parameter indicated in the superscript. We then reparameterize the distribution of $d_i^{(\cdot)}$ in terms of the effect size $\delta^{(\cdot)}$ by setting $\mu_d^{(\cdot)} = \delta^{(\cdot)}\sigma_d^{(\cdot)}$.
- Let $\lambda^{(\cdot)} := \frac{\sigma_d^{(\cdot)}}{\sigma_m^{(\cdot)}}$ be the ratio of the standard deviation of the effect to the standard deviation of the mean. We then reparameterize the distribution of $d_i^{(\cdot)}$ in terms of $\lambda^{(\cdot)}$ by setting $\sigma_d^{(\cdot)} = \lambda^{(\cdot)}\sigma_m^{(\cdot)}$.

Using this parameterization, we can rewrite the distribution of $\vec{d}_i$ as

$$\begin{pmatrix} d_i^{\alpha+} \\ d_i^{\alpha-} \\ d_i^{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta^{\alpha+}\lambda^{\alpha+}\sigma_m^{\alpha+} \\ \delta^{\alpha-}\lambda^{\alpha-}\sigma_m^{\alpha-} \\ \delta^{\beta}\lambda^{\beta}\sigma_m^{\beta} \end{pmatrix}, \text{(HM2)} \right.$$
$$\left. \begin{pmatrix} \lambda^{\alpha+}\sigma_m^{\alpha+} & 0 & 0 \\ 0 & \lambda^{\alpha-}\sigma_m^{\alpha-} & 0 \\ 0 & 0 & \lambda^{\beta}\sigma_m^{\beta} \end{pmatrix} R_d \begin{pmatrix} \lambda^{\alpha+}\sigma_m^{\alpha+} & 0 & 0 \\ 0 & \lambda^{\alpha-}\sigma_m^{\alpha-} & 0 \\ 0 & 0 & \lambda^{\beta}\sigma_m^{\beta} \end{pmatrix} \right).$$

The following observations motivate this parameterization:

- Inference on the effects of the experimental factor on some parameter can be performed by comparing a model where the distribution of the effect is restricted to a model without this restriction, by means of the ratio of the marginal model likelihoods (i.e., the Bayes factor, Kass & Raftery, 1995). In particular, restricting $\delta^{(\cdot)}$ to 0 provides a test of the fixed effect of the factor on the respective parameter, and restricting $\lambda^{(\cdot)}$ to 0 provides a test of the random effect.
- The parameterization of the standard deviation of the random effect $\sigma_d^{(\cdot)}$ in terms of the ratio $\lambda^{(\cdot)}$ allows the formulation of informed priors for the random effects without precise prior knowledge of the variance of participant-level

parameters. This greatly facilitates the formulation of priors that are informed enough for valid Bayesian inference (Vanpaemel, 2010). To explain this idea in more detail: Without prior knowledge of the standard deviation of the parameter in the population $\sigma_m^{(\cdot)}$, it is difficult to define a plausible range for the standard deviation of the random effect of the experimental factor $\sigma_d^{(\cdot)}$. In contrast, it is intuitive to make assumptions about the ratio of the standard deviation of the random effect with regards to the standard deviation of the parameter $\lambda^{(\cdot)}$. In particular, we deem it plausible that the variance of the random effect is considerably smaller than the variance of the parameter (so $0 \leq \lambda^{(\cdot)} < 1$).

- Similarly, it is difficult to predefine a range of meaningful values for experimental fixed effects without prior knowledge of the random effect (i.e., how strongly parameters can be expected to vary unsystematically between factor levels due to differences between observatory units/participants). However, it is feasible to make informed assumptions about the distribution of $\delta^{(\cdot)}$, which puts the size of the fixed effect on the scale of its associated random effect (for similar approaches in the context of Bayesian t-tests see Gönen et al., 2005; Rouder et al., 2009).

Priors

Using the parameterization described in section 2 (**HM1** and **HM2**), we need to formulate priors for the population-level parameters defining the distribution of participant-level parameter means ($\mu_m^{\alpha+}$, $\mu_m^{\alpha-}$, μ_m^{β}, $\sigma_m^{\alpha+}$, $\sigma_m^{\alpha-}$, σ_m^{β} and R_m), as well as the parameters defining the distribution of experimental effects ($\delta^{\alpha+}$, $\delta^{\alpha-}$, δ^{β}, $\lambda^{\alpha+}$, $\lambda^{\alpha-}$, λ^{β} and R_d). As our research interest is not focussed on inference about the parameter means, we formulate the weakly informative priors

$$\begin{aligned} \mu_m^{\alpha+}, \mu_m^{\alpha-}, \mu_m^{\beta} &\sim \mathcal{N}(0,1), \\ \sigma_m^{\alpha+}, \sigma_m^{\alpha-}, \sigma_m^{\beta} &\sim \text{Halfnormal}(1), \\ R_m &\sim \text{LKJ}(2), \end{aligned}$$

where Halfnormal(1) denotes the halfnormal (or folded normal) distribution with standard deviation 1, and LKJ(2) denotes the LKJ distribution on the space of correlation matrices with the hyperparameter 2 (Lewandowski et al., 2009, see also https://mc-stan.org/docs/2_22/functions-reference/lkj-correlation.html).

As explained with regards to the parameterization of experimental effects (**HM2**), models without a fixed or random effect on a certain parameter may be obtained by restricting $\delta^{(\cdot)}$ or $\lambda^{(\cdot)}$ to 0, respectively. At the same time, models that do allow for fixed or random effects must be equipped with informed priors, to allow for valid inference. For the size of the fixed effects, we thus formulate the prior

$$\delta^{(\cdot)} \sim \mathcal{N}(0, \epsilon),$$

where $\epsilon = 0$ for models without the respective fixed effect (resulting in a point-mass prior of $\delta^{(\cdot)}$ on 0), and $\epsilon > 0$ for models in which the fixed effect may be different from zero. For models with fixed effect we choose $\epsilon = 1$, resulting in a standard normal prior which puts the majority of probability mass on effects with a magnitude smaller than 1, but also allows for greater effects. Similarly, we define the prior on

the ratio of standard deviations as

$$\lambda^{(\cdot)} \sim \text{Beta}(1, \eta),$$

where $\eta = \infty$ for models without the respective random effect (resulting in a point-mass prior of $\lambda^{(\cdot)}$ on 0), and $\eta < \infty$ for models in which the standard deviation of the random effect is allowed to be greater than zero. For models with random effect we choose $\eta = 2$, reflecting our assumption that the standard deviation of the random effects should be considerably smaller than the standard deviation of the parameter. We complete the prior specification by putting a weakly informed prior on the correlation matrix R_d,

$$R_d \sim \text{LKJ}(2).$$

5.B Supplementary Tables

Here, we provide the Supplementary Tables 5.B.1 and 5.B.2.

TABLE 5.B.1: Brain activity during the outcome phase. We report the first local maximum within each cluster. Effects were tested for significance using cluster inference with a cluster defining threshold of $p < 0.001$ and a cluster probability of $p < 0.05$, FWE-corrected. * For these effects, we list the 10 clusters with the highest peak z-values.

	MNI Coordinates				
Contrast and Brain Region	**x**	**y**	**z**	**z-value**	**Cluster size**
Outcome phase					
Positive >Negative *					
R caudate	12	10	-12	>10	276
L caudate	-10	10	-14	7.63	284
L paracentral lobule	-4	-28	56	7.00	4920
L superior frontal gyrus	-14	24	54	5.64	912
R middle frontal gyrus	28	20	54	5.41	1092
R superior temporal gyrus	66	-14	-2	5.17	274
R middle occipital gyrus	54	-68	26	4.98	555
L/R superior frontal gyrus, medial orbital	-2	50	-12	4.90	448
L inferior temporal gyrus	-60	-4	-28	4.83	175
L angular gyrus	-46	-66	34	4.73	398
Negative >Positive *					
R insula	38	22	4	>10	5044
L insula	-30	22	6	>10	2157
R supplementary motor area	4	6	60	>10	3434
L Cerebellum	-30	-58	-28	>10	619
R middle frontal gyrus	46	0	54	7.18	911
R caudate	10	8	4	7.03	102
L supramarginal gyrus	-56	-40	24	6.93	1453
L thalamus	-4	-24	-4	6.55	697
R cerebellum	44	-54	-36	6.49	366
$(\text{Positive - Negative})_{\text{Self}}$ >$(\text{Positive - Negative})_{\text{Other}}$					
L/R paracentral lobule	0	-32	60	6.54	883
L postcentral gyrus	-36	-26	52	5.96	1217
R postcentral gyrus	64	-10	30	4.50	73
L precuneus	-12	-54	36	4.29	181
R angular gyrus	52	-66	28	4.13	195
L middle temporal gyrus	-46	-54	20	4.11	76
R middle frontal gyrus	28	18	54	4.06	262
R calcarine fissure	4	-60	10	4.00	63
R middle frontal gyrus	42	18	42	3.97	62
$(\text{Negative - Positive})_{\text{Self}}$ >$(\text{Negative - Positive})_{\text{Other}}$					
R rolandic operculum	56	2	6	>10	1997
L/R middle cingulate cortex	6	4	36	6.97	1936
L superior temporal pole	-54	4	0	6.72	716
L supramarginal gyrus	-62	-24	18	5.59	419
R postcentral gyrus	22	-46	62	5.51	397
L insula	-32	26	6	5.11	71
R brainstem	8	-20	-14	4.65	85
L Cerebellum	-26	-62	-26	4.65	119

TABLE 5.B.2: Parametric modulation during the outcome phase. |PE| = absolute prediction error. Effects were tested by calculating paired-sample t-tests on contrasts of the first-level parameter estimates of the parametric modulator. In the following, |PE|Self,Positive denotes the parameter estimate of the parametric modulator of the absolute prediction error for positive outcomes in the Self condition. The other terms are defined correspondingly. We report the first local maximum within each cluster. Effects were tested for significance using cluster inference with a cluster defining threshold of $p < 0.001$ and a cluster probability of $p < 0.05$, FWE-corrected.

$$|PE|_{Mean} = (|PE|_{Self,Positive} + |PE|_{Self,Negative} + |PE|_{Other,Positive} + |PE|_{Other,Negative})/4$$
$$|PE|_{Positive} = (|PE|_{Self,Positive} + |PE|_{Other,Positive})/2$$
$$|PE|_{Negative} = (|PE|_{Self,Negative} + |PE|_{Other,Negative})/2$$
$$|PE|_{Self} = (|PE|_{Self,Positive} + |PE|_{Self,Negative})/2$$
$$|PE|_{Other} = (|PE|_{Other,Positive} + |PE|_{Other,Negative})/2$$

	MNI Coordinates				
Contrast and Brain Region	**x**	**y**	**z**	**z-value**	**Cluster size**
Parametric modulator: absolute prediction error					
$\|PE\|_{Mean} > 0$					
L inferior frontal gyrus, triangular part	-48	36	20	5.59	949
L inferior parietal lobule	-34	-48	36	5.53	1548
R precentral gyrus	48	8	32	5.33	3913
R inferior parietal lobule	36	-54	46	5.24	2786
L precentral gyrus	-38	0	28	4.88	666
R inferior temporal gyrus	44	-34	-30	4.20	241
R insula	30	26	-4	4.13	87
$\|PE\|_{Positive} > \|PE\|_{Negative}$					
R inferior parietal lobule	44	-54	50	7.40	3163
L Cerebellum	-28	-62	-36	7.11	1340
R middle frontal gyrus	38	48	24	6.57	6670
L middle frontal gyrus	-44	30	34	5.96	1333
R middle cingulate gyrus	2	32	34	5.87	1004
R thalamus/striatum	14	2	16	5.39	698
L inferior parietal lobule	-26	-60	40	4.98	66
L inferior parietal lobule	-36	-46	34	4.95	1017
R inferior temporal gyrus	64	-32	-12	4.94	665
R precuneus	8	-66	42	4.60	204
$\|PE\|_{Negative} > \|PE\|_{Positive}$					
L/R fusiform gyrus	-26	-50	-14	6.99	9236
L/R postcentral gyrus	58	-16	50	6.64	12189
L/R superior frontal gyrus, medial orbital	-10	42	-12	5.41	533
L/R olfactory cortex	-4	16	-14	4.74	87
R parahippocampal gyrus	22	-12	-26	4.45	86
$\|PE\|_{Other} > \|PE\|_{Self}$					
R Cerebellum	32	-70	-32	4.44	194
L middle temporal gyrus	-66	-40	-14	4.39	75
R superior frontal gyrus	24	62	16	4.31	238
L inferior frontal gyrus, orbital	-38	28	-16	4.24	87
R precentral gyrus	36	-18	52	4.16	113
R inferior frontal gyrus, orbital	34	32	-20	4.11	78
L superior frontal gyrus	-10	50	26	4.11	271
L middle temporal gyrus	-56	-30	0	3.97	66
R superior parietal gyrus	24	-66	52	3.82	76

Chapter 6

Sampling in Constrained Space: Efficient Estimation of Model Evidence under Equality and Inequality Constraints

Abstract

Many scientific theories imply equality and inequality constraints on the statistical model underlying the observed data. Incorporating such constraints into hypothesis testing increases the efficiency and precision of inference. However, existing tools of testing such hypotheses are computationally expensive and do not scale well to complex sets of constraints. Here, we introduce Sampling in Constrained Space (SICS), a highly general approach for performing Bayesian hypothesis testing on constrained models. Through a simple reparameterization, SICS allows the efficient sampling of the posterior distribution from those sections of the parameter space that fulfill all constraints. This, in turn, enables estimating the marginal likelihood under the constrained model, through established algorithms such as bridge sampling. We demonstrate the versatility of SICS by analyzing a learning task dataset using constrained reinforcement learning models, from both a single-level and hierarchical perspective. Our results show that SICS outperforms previous approaches in terms of precision and efficiency. By providing guidelines on its application, we hope to make SICS an integral part of scientists' inferential toolkit.

6.1 Introduction

When we perform inference, we assess how well empirical observations are explained by statistical models. In many cases, scientific expectations can be translated to constraints on the model's parameters. For example, researchers conducting a clinical trial of a new antidepressant might expect the treatment group to show, on average, less severe depressive symptoms than the placebo group, which in turn shows less severe symptoms than the control group. Similar constraints appear in many domains whenever theories imply directional or structural expectations. Successful applications of constrained statistical modeling range from analysis of variance (ANOVA)

This chapter is currently under review by the *Journal of Mathematical Psychology*, and preprinted as **Lengersdorff, L. L.**, & Marsman, M. (2025). Sampling in Constrained Space: Efficient Estimation of Model Evidence under Equality and Inequality Constraints. *PsyArXiv*. https://doi.org/10.31234/osf.io/a3ytg_v1

(Klugkist, Laudy, & Hoijtink, 2005) and correlation analysis (Mulder, 2016) to structural equation modeling(Gu et al., 2019) and meta-analysis (Haaf & Rouder, 2023). Whenever theories impose constraints, it is in the researchers' interest to incorporate them as precisely as possible into their statistical tests. Not only does this ensure a high congruence between theory and model, but tightly constrained hypotheses are also highly informative, increasing the efficiency and power of inferential tests (van de Schoot & Strohmeier, 2011). As a result, the testing of constrained hypotheses has become an intensively researched topic in statistical inference (Hoijtink, 2011; Klugkist, Kato, & Hoijtink, 2005; Mulder et al., 2022), with specialized tests and software being developed for a wide variety of models (e.g., Mulder et al., 2021).

Most methods for constrained hypotheses are based on Bayesian model comparison, where evidence for a hypothesis is quantified as the marginal likelihood of the observed data under the assumption that the hypothesis is true. However, estimating this likelihood is a non-trivial problem, as it typically involves high-dimensional integrals that are intractable in closed form. For constrained hypotheses, marginal likelihood estimation is even more complicated. Constraints involving multiple parameters result in a statistical model that is only defined on a small and irregular portion of the parameter space. This leads to an irregular integration domain that is difficult to handle by traditional numerical integration approaches.

Several methods have been developed to estimate the evidence for constrained hypotheses despite these challenges. One of the most widely used techniques is the encompassing prior (EP) approach (Klugkist, Kato, & Hoijtink, 2005). In the *unconditional* EP method, samples are drawn - typically via Markov chain Monte Carlo (MCMC) methods (Robert et al., 1999) - from the unconstrained, or encompassing, prior and posterior distributions of the parameter. Evidence for the constrained hypothesis is then estimated, using Monte Carlo integration, as the ratio of posterior samples that satisfy the constraint to prior samples that satisfy the constraint; in essence, it assesses how much the data shift the probability mass into the region of the parameter space that satisfies the constraint. While elegant and straightforward, this approach becomes highly inefficient when the constraint defines only a small subregion of the parameter space, or when the posterior density is bound away from the constrained parameter space. In such cases, only a tiny fraction of the samples fall within the constrained region, making the estimate highly unstable unless an impractically large number of samples are drawn. The *conditional* EP method (Mulder et al., 2010) mitigates this problem by decomposing the constraint into partial sub-constraints and estimating the evidence for each one sequentially. This stabilizes the final estimate, but at the cost of significantly increased computational complexity. As the number of sub-constraints increases, so does the runtime of the estimation - meaning that the conditional EP method does not scale well to multi-parameter constrained hypotheses. In addition, the precision of the final estimate may degrade if any of the intermediate steps are unreliable.

In this paper, we present *Sampling in Constrained Space* (SICS), an efficient and scalable estimation approach designed to overcome the limitations of existing methods. The core idea behind SICS is simple: instead of sampling from an unconstrained posterior and then checking whether the samples satisfy the constraint, we redefine the statistical model so that the constraint is automatically satisfied while sampling from the model. This is achieved by a reparameterization that transforms the original model into one where the parameter is guaranteed to lie within the constrained region. By

building the constraint directly into the structure of the model, SICS avoids the inefficiencies of Monte Carlo integration and the complexity of decomposing constraints. This makes the method computationally efficient, even when dealing with complex or highly restrictive hypotheses. Once the model is reparameterized, the marginal likelihood of the constrained hypothesis can be estimated using modern techniques such as bridge sampling (Gronau et al., 2017), allowing for accurate and scalable Bayesian testing of constrained hypotheses.

The rest of the paper is structured as follows: Section 2 introduces the theory behind Bayesian testing of constrained hypotheses and provides an overview of encompassing prior methods, including a discussion of their limitations. Section 3 presents the SICS method, beginning with intuitive explanations and then stating and proving its central lemmas. Section 4 demonstrates the flexibility and efficiency of SICS in an empirical example, where we apply it to the testing of constrained hypotheses in both single-subject and multilevel reinforcement learning models. Section 5 concludes the paper with an outlook on possible further developments of SICS, in particular with respect to software implementations.

6.2 Bayesian Testing of Constrained Hypotheses

Here, we lay out the mathematical foundations of Bayesian hypothesis testing with parameter constraints. In Section 2.1, we introduce the general framework of testing with Bayes factors, including the role of priors, posteriors, and marginal likelihoods. We then turn to the specific challenges posed by constrained hypotheses and explain how such hypotheses can be represented by conditioning the statistical model on a constraint. In Section 2.2, we briefly discuss the encompassing prior method as the most commonly used solution strategy, along with its limitations.

6.2.1 Mathematical Background

One of the most established approaches to testing constrained hypotheses is Bayesian hypothesis testing using the *Bayes Factor* (BF; Dienes, 2016; Kass & Raftery, 1995). The BF of a hypothesis $\mathcal{H}_i$ against another hypothesis $\mathcal{H}_j$ is defined as the ratio

$$\mathrm{BF}_{ij} = \frac{p(D \mid \mathcal{H}_i)}{p(D \mid \mathcal{H}_j)},$$

where $p(D \mid \mathcal{H}_i)$ is the *marginal likelihood* of the observed data D under $\mathcal{H}_i$[1]. It is given by the integral

$$p(D \mid \mathcal{H}_i) = \int_\Theta p(D \mid \theta)\, \pi(\theta \mid \mathcal{H}_i)\, \mathrm{d}\lambda(\theta) \tag{6.1}$$

where θ is a vector of k parameters $(\theta_1, \theta_2, ..., \theta_k)$; $\Theta \subseteq \mathbb{R}^k$ is the k-dimensional *parameter space*, the set of all possible values that θ may take; $p(D \mid \theta)$ is the likelihood

[1] A note on mathematical notation: In this paper, we will use the lowercase letter p to denote the likelihood function and the marginal likelihood, and the Greek letter π to denote priors, posteriors, and integrals over these functions, such as normalizing constants. We will also use the bold blackboard letter $\mathbb{P}$, as in $\mathbb{P}(\theta \in C)$, to emphasize that a term is the probability of a random event. Contrast this with terms such as the marginal likelihood $p(D)$, which cannot be interpreted as a probability if D is a continuous variable. Also, we will use the floating dot $\cdot$ as a placeholder, as in $\pi(\cdot \mid D)$, when we want to emphasize that we are referring to the function itself, and not to its evaluation at a particular argument.

function; $\pi(\theta \mid \mathcal{H}_i)$ is the density of the *prior* of θ under $\mathcal{H}_i$, and λ is the k-dimensional Lebesgue measure. The marginal likelihood quantifies how well the data are predicted by the model, averaged over all plausible parameter values.

Formulating a constrained hypothesis $\mathcal{H}_C$ can now be understood as *conditioning* the model on the event that the parameter θ lies in a subset C of the parameter space. Using a term from constrained optimization, we will refer to C as the *feasible set* of $\mathcal{H}_C$. Because the feasible set fully determines the hypothesis, we denote constrained hypotheses with their feasible set as a subscript, i.e. $\mathcal{H}_C$ is the constrained hypothesis with the feasible set C.

From a statistical point of view, satisfying the constraint $\theta \in C$ can be treated as an event that is certain if θ lies in C, and impossible if θ is outside of C. The "likelihood function" associated with $\theta \in C$ is thus simply

$$\pi(\theta \in C \mid \theta) = \mathbb{I}_C(\theta) = \begin{cases} 0, & \theta \notin C \\ 1, & \theta \in C \end{cases} \tag{6.2}$$

where $\mathbb{I}_C$ is the indicator function of the feasible set C. Treating the constraint satisfaction as we would normally treat observed data, we can now apply Bayes' theorem to obtain the *constrained prior* of θ given that $\theta \in C$:

$$\begin{aligned} \pi(\theta \mid \theta \in C) &= \frac{\mathbb{P}(\theta \in C \mid \theta)\, \pi(\theta)}{\pi(\theta \in C)} \\ &= \mathbb{I}_C(\theta) \cdot \frac{\pi(\theta)}{\pi(\theta \in C)}. \end{aligned}$$

Intuitively, the constrained prior $\pi(\theta \mid \theta \in C)$ encodes our beliefs about possible values of θ when we additionally assume that θ satisfies the constraint. The term in the denominator $\pi(\theta \in C)$ is the *normalizing constant* of the constrained prior,

$$\pi(\theta \in C) = \int_C \pi(\theta)\, \mathrm{d}\lambda_C(\theta), \tag{6.3}$$

where λ_C is a (possibly lower-dimensional) version of the Lebesgue measure concentrated on C.[2] The normalizing constant $\pi(\theta \in C)$ is the unique constant such that integrating the conditional prior $\pi(\theta \mid \theta \in C)$ over C yields 1. When C is of full dimension, $\pi(\theta \in C)$ is equal to the prior probability that the constraint is satisfied. For lower-dimensional C it becomes a marginal density. This is the case when the tested constraint contains a strict equality. For example, in a two-dimensional parameter space, the constraint $\theta_1 = \theta_2$ reduces the feasible set to the one-dimensional line where θ_1 is exactly equal to θ_2.

In principle, the prior π can be chosen in any way that reflects the researcher's assumptions and knowledge about θ. However, for testing a constrained hypothesis $\mathcal{H}_C$ against alternatives, the prior should assign plausible and consistent probabilities to values of θ both inside and outside of C. Since such a prior has support over the entire parameter space, it is also often called an *encompassing prior* (Klugkist, Kato, & Hoijtink, 2005).

[2] The measure λ_C is defined as follows: Let k be the dimension of Θ, and l the dimension of C. Further, let $\lambda_{\mathbb{R}^l}$ be the Lebesgue measure on $\mathbb{R}^l$, and T be a function that bijectively maps $\mathbb{R}^l$ to C. Then $\lambda_C := \lambda_l \circ T^{-1}$. A possible form of T is given in Lemma 1 of this paper.

In the same way as above, we obtain the *constrained posterior* $\pi(\theta \mid D, \theta \in C)$ by conditioning the unconstrained posterior $\pi(\theta \mid D)$ on the constraint $\theta \in C$. Equivalently, we can understand it as the result of conditioning the constrained prior $\pi(\theta \mid \theta \in C)$ on the event that data D was observed. Either way, we get

$$\pi(\theta \mid D,\ \theta \in C) = \frac{p(D \mid \theta)\, \pi(\theta \mid \theta \in C)}{p(D \mid \theta \in C)}, \tag{6.4}$$

where $p(D \mid \theta \in C)$ is the marginal likelihood of the data under the constrained model,

$$p(D \mid \theta \in C) = \int_C p(D \mid \theta)\, \pi(\theta \mid \theta \in C)\, \mathrm{d}\lambda_C(\theta). \tag{6.5}$$

This constant is crucial in Bayesian hypothesis testing, since the BF is the ratio of the marginal likelihoods of two competing hypotheses. Its efficient and precise estimation is therefore important.

6.2.2 Estimating Model Evidence under Constraints: Existing Approaches

Calculating the marginal likelihood is a notoriously difficult problem. In most cases, integral (6.1) has no analytic solution, and must be approximated by numerical methods (Chib & Jeliazkov, 2001; Wasserman, 2000). Constrained hypotheses add another layer of complexity, since their feasible sets often form complex integration regions. The *encompassing prior* (EP) methods (Klugkist, Kato, & Hoijtink, 2005) avoid these complications by not estimating the marginal likelihood of a constrained hypothesis $p(D \mid \theta \in C)$ itself. Instead, they estimate the BF comparing the constrained hypothesis $\mathcal{H}_C$ to the unconstrained, or *encompassing*, hypothesis $\mathcal{H}_\Theta$:

$$\mathrm{BF}_{C\Theta} = \frac{p(D \mid \theta \in C)}{p(D \mid \theta \in \Theta)} = \frac{p(D \mid \theta \in C)}{p(D)}.$$

The EP methods make use of the following result: By viewing $\theta \in C$ as a random event and applying Bayes' theorem, we can express the marginal likelihood of the constrained hypothesis as:

$$p(D \mid \theta \in C) = \frac{\pi(\theta \in C \mid D)\, p(D)}{\pi(\theta \in C)}.$$

Dividing both sides of the equation by $p(D)$, we obtain

$$\frac{p(D \mid \theta \in C)}{p(D)} = \mathrm{BF}_{C\Theta} = \frac{\pi(\theta \in C \mid D)}{\pi(\theta \in C)}. \tag{6.6}$$

This relation suggests the particularly intuitive interpretation of $\mathrm{BF}_{C\Theta}$ as the ratio of the posterior to the prior probability that the constraint is satisfied (i.e., how much the data shift the probability mass into or out of the feasible set). The EP method now uses the fact that it is easy to estimate this ratio using Monte Carlo methods.

The Unconditional Encompassing Prior Method

In the unconditional EP method, $\mathrm{BF}_{C\Theta}$ is estimated by Monte Carlo integration, with variations depending on whether the constraint being tested contains strict equalities.

For pure inequality constraints, the prior and posterior support for these constraints $\pi(\theta \in C)$ and $\pi(\theta \in C \mid D)$ are estimated by sampling from the prior and posterior distributions and computing the proportion of samples that satisfy the constraint. Given N_0 prior samples $\hat{\theta}_i \sim \pi$ and N_1 posterior samples $\tilde{\theta}_i \sim \pi(\cdot \mid D)$, by the law of large numbers we have:

$$\text{BF}_{C\Theta} = \frac{\mathbb{P}(\theta \in C \mid D)}{\mathbb{P}(\theta \in C)} \approx \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} \mathbb{I}(\tilde{\theta}_i \in C)}{\frac{1}{N_0}\sum_{i=1}^{N_0} \mathbb{I}(\hat{\theta}_i \in C)}. \tag{6.7}$$

Such samples can be generated using standard MCMC algorithms, such as Hamiltonian Monte Carlo (Betancourt, 2017). For constrained hypotheses that include equality constraints, the feasible set becomes lower-dimensional, and additional care must be taken when computing marginal likelihoods. In such cases, adaptations of the encompassing prior method have been proposed, e.g. by decomposing the BF into components for the equality and inequality parts (cf. Sarafoglou et al., 2023).

The unconditional EP method has gained popularity due to its simplicity (Haaf & Rouder, 2017, 2023; Haaf et al., 2020). However, it performs very poorly when the prior resp. posterior have very little overlap with the feasible set - which is often the case when many constraints are put onto a high-dimensional parameter. Then, the probability of an MCMC sample satisfying all constraints is very low, which is detrimental to the precision of the BF.

Consider, for example, an ordinal constraint $\theta_1 < \theta_2 < \cdots < \theta_k$ under an uninformative prior. In this case, it can be shown that $\mathbb{P}(\theta \in C) = 1/k!$, which decreases super-exponentially with k. For $k = 10$, this yields $\mathbb{P}(\theta \in C) = 1/3628800$, which means that on average you would have to draw 3.63 million prior samples to get one observation that satisfies the constraint. The situation may be even be worse for the posterior: if the constraint is not well supported by the data, most of the posterior mass may lie outside C, further reducing $\mathbb{P}(\theta \in C \mid D)$.

The relative error of the Monte Carlo probability estimate is inversely proportional to the probability itself (see Appendix 6.A for mathematical details). Thus, if $\mathbb{P}(\theta \in C)$ or $\mathbb{P}(\theta \in C \mid D)$ are close to zero, an impractically large number of samples are required to obtain an estimate with an acceptable level of precision. Continuing our earlier example, estimating $\mathbb{P}(\theta \in C) = \frac{1}{3{,}628{,}800}$ with a relative error of less than 10% would require over 362 million samples — an enormous computational cost, even with efficient samplers.

Thus, while the unconditional EP method is computationally and conceptually simple, using it to estimate a highly constrained hypothesis with enough precision is often infeasible.

The Conditional Encompassing Prior Method

The conditional encompassing method was developed to enable the estimation of BFs for hypotheses that contain many constraints (Hoijtink, 2011; Mulder et al., 2010). Its central idea is to factorize the BF into partial Bayes factors, which are then estimated independently by Monte Carlo integration. Let C be the feasible set of a constrained hypothesis consisting of k inequality constraints. Further, let C_i be the feasible set of

the i-th constraint. Then the BF can be written as

$$\begin{aligned} \text{BF}_{C\Theta} &= \frac{\mathbb{P}\Big((\theta \in C_1) \cap (\theta \in C_2) \cap \ldots \cap (\theta \in C_k) \mid D\Big)}{\mathbb{P}\Big((\theta \in C_1) \cap (\theta \in C_2) \cap \ldots \cap (\theta \in C_k)\Big)} \\ &= \underbrace{\frac{\mathbb{P}(\theta \in C_1 \mid D)}{\mathbb{P}(\theta \in C_1)}}_{\text{BF}_{\text{partial 1}}} \cdot \underbrace{\frac{\mathbb{P}(\theta \in C_2 \mid \theta \in C_1, D)}{\mathbb{P}(\theta \in C_2 \mid \theta \in C_1)}}_{\text{BF}_{\text{partial 2}}} \cdot \ldots \cdot \underbrace{\frac{\mathbb{P}(\theta \in C_k \mid \theta \in \bigcap_{i=1}^{k-1} C_i, D)}{\mathbb{P}(\theta \in C_k \mid \theta \in \bigcap_{i=1}^{k-1} C_i)}}_{\text{BF}_{\text{partial k}}} \end{aligned} \tag{6.8}$$

where $\bigcap_{i=1}^{k-1} C_i$ is the intersection of the sets C_1 to C_{k-1}. Thus, the BF is factored into the product $\prod_{i=1}^{k} \text{BF}_{\text{partial i}}$, where $\text{BF}_{\text{partial i}}$ is the BF of the i-th constraint in the encompassing model *conditional* on the fulfillment of all constraints up to $i-1$. In conditional EP, each of these k BFs is estimated using Monte Carlo integration as in (6.7), with samples taken from the corresponding conditional prior or posterior distribution.

For tightly constrained hypotheses, the conditional EP method can achieve much higher precision than the unconditional EP method. This is because the probabilities $\mathbb{P}(\theta \in C_i \mid \theta \in C_1, \ldots, \theta \in C_{i-1})$ are bound to be much larger than the probability of the full constraint $\mathbb{P}(\ \in \mathbb{C})$. For the interested reader, we derive the relative error of the conditional EP method in Appendix 6.A. Returning to our earlier example, estimating $\mathbb{P}(\theta \in C) = \frac{1}{3{,}628{,}800}$ with a relative error of 10% can be achieved using the conditional EP method with a sample of only 40,670 - a substantial improvement over the 362 million samples required by the unconditional EP method.

The improved precision and efficiency of the conditional EP has led to its wide adoption and implementation in several statistical software packages (Gu et al., 2019; Heck & Davis-Stober, 2019; Mulder, 2016; Mulder et al., 2012, 2021). However, it still suffers from limitations. For one, the precision of the estimate may still deteriorate if one of the partial feasible sets C_i is small, or if the posterior distribution is strongly concentrated outside of C_i. Then, the corresponding partial BF will also be estimated with low precision, which will affect the precision of the overall estimate. Moreover, the computational cost of the conditional EP scales linearly with the number of constraints. Each additional constraint requires the computation of an additional partial BF, for which sufficiently large prior and posterior samples must be drawn.

As we have seen, the main limitations of the EP methods stem from their reliance on Monte Carlo integration, which becomes increasingly inefficient as the constraints become tighter or more numerous. These issues can be overcome by avoiding sampling from the full parameter space - that is, by ensuring that all samples inherently satisfy the constraint. This is the central idea behind Sampling in Constrained Space (SICS), which we introduce in the next section.

6.3 Sampling in constrained space

This section is organized as follows. Section 3.1 introduces the central idea of SICS in an intuitive way. Section 3.2 illustrates the method by going through a step-by-step solution of a simple example problem. Section 3.3 fully generalizes the approach and provides the key lemmas and proofs. Section 3.4 concludes the theoretical part of the manuscript with a brief elaboration on the method's precision.

6.3.1 Conceptual Overview

The first key step of SICS is constrained sampling: Instead of *sampling first from the posterior first then discarding samples that fall outside the feasible set* ("sample, then constrain"), we *first constrain the parameter space to the feasible set and then sample from the constrained posterior* ("constrain, then sample"), ensuring that every sample automatically satisfies the constraint. The second key step is that, once we have samples from the constrained posterior, we can estimate the marginal likelihood of the constrained hypothesis using standard techniques such as bridge sampling (Gronau et al., 2017).

Constrained sampling is possible because conditioning is *commutative*: When we go from the unconstrained prior $\pi(\theta)$ to the constrained posterior $\pi(\theta \mid \theta \in C, D)$, it does not matter if we first condition on the data D and then on the constraint $\theta \in C$, or vice versa. This idea is illustrated in Figure 6.1. The path from the prior to the posterior (A $\rightarrow$ B) represents the standard application of Bayes' theorem: the prior is updated to the posterior by conditioning on the data. There are now two different approaches of incorporating the constraint:

- **Sample, then constrain (A $\rightarrow$ B $\rightarrow$ D):** One approach is to first obtain the posterior and then apply the constraint (B $\rightarrow$ D), resulting in the constrained posterior. This is the route taken by EP methods, where conditioning on the constraint is achieved by discarding samples that fall outside the feasible set.
- **Constrain, then sample (A $\rightarrow$ C $\rightarrow$ D):** In contrast, SICS first conditions on the constraint before incorporating data. Instead of discarding samples, SICS reparameterizes the model so that the parameter θ is expressed as a function of a new parameter ω that is automatically constrained. This reparameterization transforms the prior into a constrained prior (A $\rightarrow$ C), where the parameter space is restricted to the feasible set. From this constrained prior, we then move to the constrained posterior (C $\rightarrow$ D) by conditioning on the data, just as in standard Bayesian updating. In practice, this is achieved by drawing MCMC samples from the posterior of the constrained model.

Note that the same two-step strategy - constrained sampling and marginal likelihood estimation via bridge sampling - was first used by Sarafoglou et al., 2023 to test constrained hypotheses in the multinomial model. Here, we generalize the same idea to all statistical models.

6.3.2 A Motivating Example

To illustrate the principle of SICS, consider the following scenario: Suppose we have a statistical model where the data D follows a distribution function $p(D \mid \theta)$ with a three-dimensional parameter θ. Suppose further that we have formulated some prior π on θ, and are able to draw samples from the prior π and the posterior $\pi(\cdot \mid D)$ using an MCMC sampler. Our goal is to evaluate the ordinal constraint

$$\theta_1 > \theta_2 > \theta_3 > 1.$$

with feasible set $C := \{\theta \in \mathbb{R}^3 : \theta_1 > \theta_2 > \theta_3 > 1\}$ against some alternative. For this, we need to estimate the marginal likelihood $p(D \mid \theta \in C)$, which is the normalizing

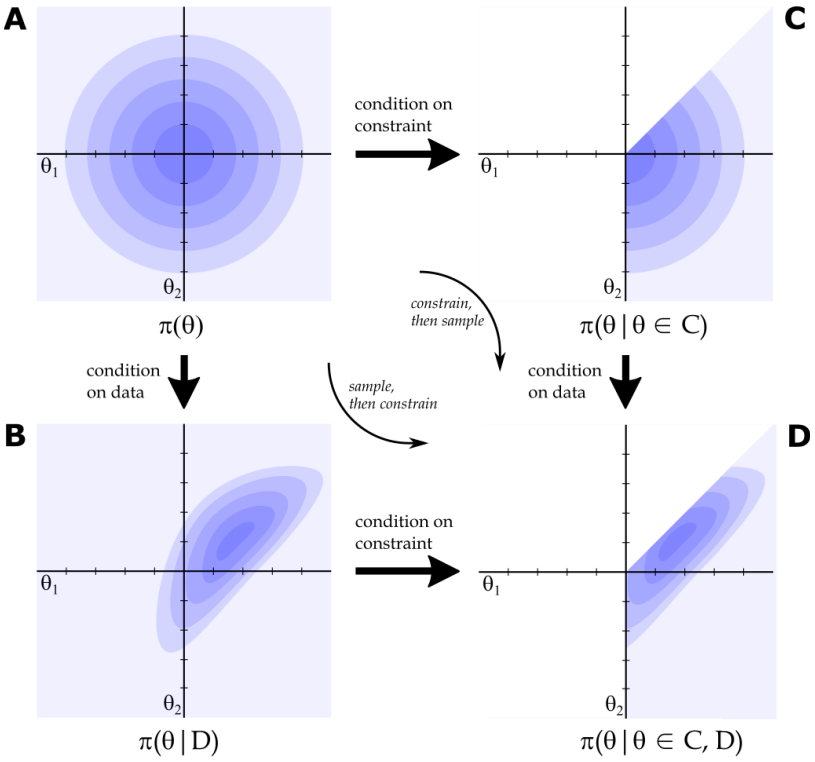

FIGURE 6.1: **Conditioning is commutative.** When deriving the constrained posterior $\pi(\theta \mid \theta \in C, D)$ from the unconstrained prior $\pi(\theta)$, the order of conditioning on data D and constraint $\theta \in C$ does not matter. The encompassing prior method follows the "sample, then constraint" route: First, $\pi(\theta)$ (panel A) is conditioned on D by drawing samples from $\pi(\theta \mid D)$ (panel B). Then, $\pi(\theta \mid D)$ is conditioned on C by discarding samples outside of C, resulting in samples from $\pi(\theta \mid \theta \in C, D)$ (panel D). In contrast, SICS follows the "constrain, then sample" route: First, the prior is conditioned on C by reparameterization. Then, the constrained prior $\pi(\theta \mid \theta \in C)$ (panel C) is conditioned on D by sampling from the constrained posterior.

constant of the constrained posterior,

$$p(D \mid \theta \in C) = \int_C p(D \mid \theta)\, \pi(\theta \mid \theta \in C)\, \mathrm{d}\lambda_C(\theta).$$

The constrained posterior is a probability distribution like any other, so in principle we should be able to estimate its normalizing constant using algorithms such as bridge sampling. However, it is not immediately clear how we can efficiently sample from this distribution. The main problem here is that the constrained posterior only has support on C, and whether or not a vector θ lies in C depends on an interaction of all three of its components. To illustrate this, consider the second component of the parameter vector, θ_2. Its support is the interval (θ_3, θ_1), so if we want to choose a valid value for θ_2, we must make sure that it lies between θ_3 and θ_1. This complicates sampling, however, because θ_3 and θ_1 are themselves random variables with restricted support.

In this simple example, we could still use an approach similar to the EP methods, where we sample from the unrestricted posterior and discard samples that do not satisfy the constraint. However, we would then be using most of our computing power to produce samples that we will ultimately discard. Moreover, as we described above, this approach will not generalize to constraints with a very small feasible set.

To solve these problems, **we need to find a way to sample from θ in such a way that the constraint is automatically satisfied**. We will achieve this by redefining the parameter vector in such a way that all of its components can vary freely and independently, while ensuring that the constraint is still satisfied. This will allow us to use standard MCMC approaches to draw samples from constrained space.

Reparameterization

We now make the following observation: the ordinal constraint $\theta_1 > \theta_2 > \theta_3 > 1$ can be decomposed into the three partial constraints

$$\theta_1 > \theta_2, \quad \theta_2 > \theta_3, \quad \text{and} \quad \theta_3 > 1,$$

which are equivalent to

$$\theta_1 - \theta_2 > 0 \quad , \quad \theta_2 - \theta_3 > 0, \quad \text{and} \quad \theta_3 - 1 > 0.$$

Thus, the constraint is satisfied whenever the three difference values

$$\begin{aligned} \omega_1 &:= \theta_1 - \theta_2, \\ \omega_2 &:= \theta_2 - \theta_3, \\ \omega_3 &:= \theta_3 - 1, \end{aligned} \tag{6.9}$$

are positive numbers. We can think of the vector $\omega := \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix}$ as a *criterion* vector: whenever every component of ω is positive, the constraint is satisfied. For example, for $\theta = \begin{pmatrix} 7 \\ 5 \\ 2 \end{pmatrix}$, $\omega = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$, and thus the constraint is satisfied. For $\theta = \begin{pmatrix} 4 \\ 5 \\ 3 \end{pmatrix}$, $\omega = \begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}$, and thus the constraint is *not* satisfied.

Note that the components of ω tell us which of the partial constraints are satisfied. For example, if we know that ω_2 is a positive number, we can be sure that $\theta_2 > \theta_3$, even without knowing the values of ω_1 and ω_3. In contrast, if we only know θ_2, we

cannot be sure that the same constraint is satisfied, without additional knowledge of θ_3. In this sense, ω provides an *orthogonal* encoding of the constraint. To know that the full constraint is satisfied, we only need to know that all of its components are positive - we do not need to know their exact values, or how those values relate to each other.

In our model, θ is a free parameter, and the criterion ω is defined as a function of θ. However, we can also *invert* this relationship: instead of specifying θ and calculating ω to assess whether the constraint is satisfied, we can specify ω and calculate the value of θ that would give rise to it. By a simple rearrangement of (6.9), we can express the components of θ as functions of the components of ω:

$$\begin{aligned} \theta_3 &= \omega_3 + 1 \\ \theta_2 &= \theta_3 + \omega_2 \\ &= \omega_1 + \omega_2 + 1 \\ \theta_1 &= \theta_2 + \omega_1 \\ &= \omega_1 + \omega_2 + \omega_3 + 1 \end{aligned} \tag{6.10}$$

So we can express θ and ω as vector-valued functions of each other:

$$\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \begin{pmatrix} \theta_1 - \theta_2 \\ \theta_2 - \theta_3 \\ \theta_3 - 1 \end{pmatrix} = \tau(\theta),$$
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} = \begin{pmatrix} \omega_1 + \omega_2 + \omega_3 + 1 \\ \omega_2 + \omega_3 + 1 \\ \omega_3 + 1 \end{pmatrix} = \tau^{-1}(\omega).$$

The function τ is a *bijective* map: we can apply it to θ to get the corresponding ω; and we can apply its inverse τ^{-1} to get θ from ω. Since τ is a *linear map*, it can also be represented using matrix notation[3]:

$$\tau(\omega) = \Gamma\theta - b, \quad \text{where } \Gamma = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$
$$\tau^{-1}(\omega) = M\omega + c, \quad \text{where } M = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Note that M is the *inverse* of Γ, $M = \Gamma^{-1}$, and $c = Mb$. Thus, τ and its inverse are fully specified by M and b:

$$\tau(\theta) = M^{-1}\theta - b, \quad \tau^{-1}(\omega) = M(\omega + b).$$

If we now choose an arbitrary positive vector ω and apply τ^{-1} to it, we are sure to get a parameter vector that satisfies the constraint. Thus we can redefine, or *reparameterize*, the model so that the constraint is always satisfied: Instead of θ, which

[3]We would like to note that we do not claim novelty. Representing equality and inequality constraints using matrix equations is standard in the mathematical field of constrained optimization, and is also already well established in constrained Bayesian hypothesis testing (e.g. Gu et al., 2019; Hoijtink, 2011; Mulder et al., 2010). Here, we only sequentially introduce these concepts to guide the reader's intuition about the necessary steps.

may lie anywhere in $\mathbb{R}^3$, the parameter of this new model is $\omega \in \mathbb{R}^3_+$, with $\tau^{-1}(\omega)$ certain to lie in the feasible set.

Next, we want to draw MCMC samples from the new parameter, as this will allow us to fit the reparameterized model to the data, and ultimately, to estimate the marginal likelihood of the constrained hypothesis. To do this, we need to find the likelihood function and the prior of the new parameter ω. Under the reparameterized model, the likelihood function is simply given by

$$p_\omega(D \mid \omega) = p(D \mid \tau^{-1}(\omega)) = p(D \mid M(\omega + b)).$$

To find the correct prior, we need to find the probability distribution on ω that is implied by our original prior on θ. To do this, we use a well-known result about the probability distribution of functions of random variables. Since ω is obtained by applying the function τ to θ, its prior density π_ω is given by

$$\pi_\omega(\omega) = \pi(\tau^{-1}(\omega) \mid \theta \in C) \cdot |D_{\tau^{-1}}(\omega)|,$$

where $\pi(\tau^{-1}(\omega) \mid \theta \in C)$ is the density function of the constrained prior evaluated at $\theta = \tau^{-1}(\omega)$, and $|(D_{\tau^{-1}}(\omega))|$ is the absolute determinant of the Jacobian matrix of τ^{-1}, evaluated at ω. Since τ^{-1} is a linear function of the form $\tau^{-1}(\omega) = M(\omega + b)$, its Jacobian is simply M. This gives us the prior

$$\begin{aligned} \pi_\omega(\omega) &= \pi(M(\omega + b) \mid \theta \in C) \cdot |M| \\ &= \frac{\pi(M(\omega + b)) \cdot |M|}{\pi(\theta \in C)} \\ &\propto \pi(M(\omega + b)) \cdot |M|, \end{aligned} \tag{6.11}$$

where $\propto$ indicates that the latter expression is proportional to the former. Using exactly the same arguments, we can also derive the posterior

$$\begin{aligned} \pi_\omega(\omega \mid D) &= \frac{p(D \mid M(\omega + b)) \cdot \frac{\pi(M(\omega+b)) \cdot |M|}{\pi(\theta \in C)}}{p(D \mid \theta \in C)} \\ &\propto p(D \mid M(\omega + b)) \cdot \pi(M(\omega + b)) \cdot |M|. \end{aligned} \tag{6.12}$$

Sampling and Marginal Likelihood Estimation

What have we achieved so far? We have managed to find a redefined model that is equivalent to the original model everywhere where the constraint holds. At the same time, the support for the new parameter ω has a much simpler form than C: we can choose arbitrary positive values for the three components of ω, and the constraint is satisfied. Contrast this with the situation for θ, where valid choices for one component depend on the values of all the other components.

It is now straightforward to draw the prior and posterior samples of ω. As seen in (6.11) and (6.12), the prior and posterior of ω are proportional to the prior and posterior of θ evaluated at $M(\omega + b)$. Thus, we need only adapt our MCMC sampler for the unconstrained model to sample from ω instead of θ, and set θ equal to $M(\omega + b)$. Once we have samples from the prior of ω, we can use an algorithm such

as bridge sampling to estimate the normalizing constant of the constrained prior,

$$\begin{aligned} K_1 &\approx \int_{\mathbb{R}^3_+} \pi(M(\omega + b)) \cdot |M| \, \mathrm{d}\lambda(\omega) \\ &= \int_C \pi(\theta) \, \mathrm{d}\lambda_C(\theta) \\ &= \pi(\theta \in C), \end{aligned}$$

where the first equality follows from the fact that $M(\cdot + b)$ is a bijective mapping between $\mathbb{R}^3_+$ and C. Similarly, we can use samples from the posterior of ω to estimate the integral

$$\begin{aligned} K_2 &\approx \int_{\mathbb{R}^3_+} p(D \mid M(\omega + b)) \, \pi(M(\omega + b)) \cdot |M| \, \mathrm{d}\lambda(\omega) \\ &= \int_C p(D \mid \theta) \, \pi(\theta) \, \mathrm{d}\lambda_C(\theta) \\ &= \pi(\theta \in C) \cdot \int_C p(D \mid \theta) \, \frac{\pi(\theta)}{\pi(\theta \in C)} \, \mathrm{d}\lambda_C(\theta) \\ &= \pi(\theta \in C) \cdot p(D \mid \theta \in C). \end{aligned}$$

Finally, we obtain the SICS estimate of the marginal likelihood as

$$p(D \mid \theta \in C) = \frac{\pi(\theta \in C) \cdot p(D \mid \theta \in C)}{\pi(\theta \in C)} \approx \frac{K_2}{K_1}.$$

As long K_1 and K_2 are consistent estimators of the two integrals involved (which is the case for bridge sampling), their ratio is a consistent estimator of $p(D \mid \theta \in C)$. Thus, we have found a way to estimate the marginal likelihood for our constrained hypothesis.

Figure 6.2 illustrates sampling from a constrained prior using a Hamiltonian Monte Carlo sampler. Note that for illustrative purposes, we have chosen to show a constraint on a two-dimensional parameter ($2\theta_2 > \theta_1 > \theta_2$) instead of the three-dimensional constraint discussed in our example. In panel B, we can clearly see that our method manages to sample exclusively from the feasible set. At the same time, the procedure maintains a sampling frequency that is proportional to the original prior. Contrast this with an unconstrained sampling procedure as illustrated in panel A, where the sampler spends a lot of time outside the feasible set. Panel C illustrates sampling within the transformed parameter space. In this example, the transformation "opens up" the parameter space along the constraint boundaries. This can also be seen in the shape of the transformed prior, which is a sheared version of the original prior within the feasible set.

6.3.3 The General Approach of Sampling in Constrained Space

Let us review the main steps we took to derive an estimator for $p(D \mid \theta \in C)$ in our example:

1. We have decomposed the full constraint into k partial constraints (in our example, $k = 3$).
2. We identified a function τ that maps the parameter vector θ onto a criterion vector ω that orthogonally encodes the partial constraints: the i-th constraint is

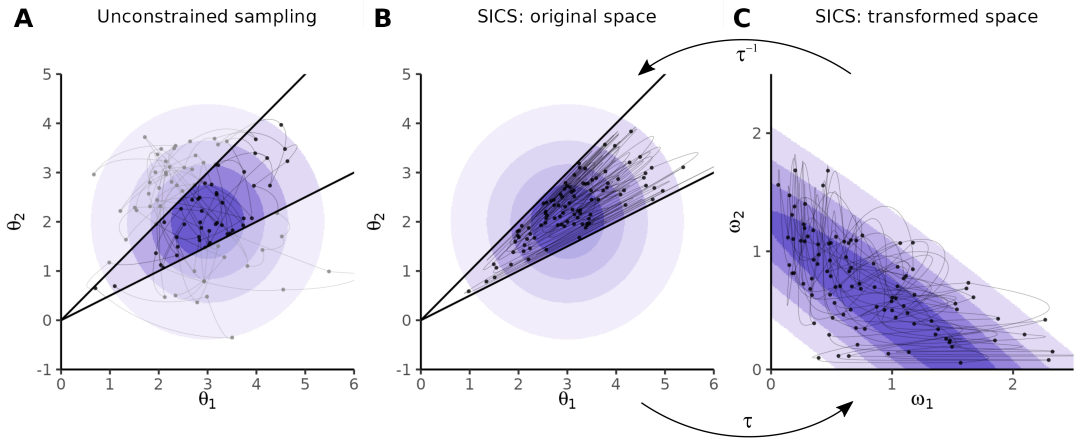

FIGURE 6.2: **Sampling in constrained Space.** Shown are samples (dots) drawn by a Hamiltonian Monte Carlo sampler and Hamiltonian trajectories (lines). **A)** Sampling from the unconstrained parameter space is inefficient because the sampler spends a lot of time in regions that do not satisfy the constraint. **B)** SICS samples exclusively within the constrained space, with probability density proportional to the encompassing prior/posterior wherever the constraint holds. **C)** The prior/posterior is transferred from the feasible set to an auxiliary space via the transformation τ. Here, MCMC sampling can proceed with standard implementations. **Notes**: Constraint: $2\theta_2 > \theta_1 > \theta_2$. $(\theta_1, \theta_2) \sim \mathcal{N}((3,2), I_2)$. Samples were drawn using naive Hamiltonian Monte Carlo sampling with $\Delta t = 0.01$ and 100 leapfrog steps. Blue contours represent regions containing, from lightest to darkest, 95%, 75%, 50%, and 25% probability mass.

satisfied if and only if the i-th component of ω is in a certain interval (in our example, all components of ω had to be positive numbers, $\omega_i \in (0, \infty)$).

3. We inverted the relationship between θ and ω, defining ω as the model's free parameter, and θ as a function of ω.
4. We sampled from the prior and posterior of ω, and used these samples to estimate the marginal likelihood.

The idea behind SICS is that these steps can be performed for any model and prior, as long as the tested constraint satisfies certain requirements. As we will see, it is necessary that the constraint can be represented by a matrix with linearly independent rows. In the following, we will introduce the necessary definitions to formalize and generalize SICS, and prove the key lemmas.

For mathematical precision, the definitions and theorems in this section are stated in more formal terms than the rest of the paper. These may be of particular interest to mathematically inclined readers. To support readers less familiar with formal notation, each definition and theorem is followed by a brief intuitive explanation of its meaning and relevance for the rest of the paper.

Linear Constraints

The constraint tested in our example was of such a form that we could perform all the necessary steps: we were able to identify a function τ that maps θ onto an orthogonal criterion vector ω, and we were able to find the inverse function τ^{-1}. This is not always possible. For example, the constraint $\theta_1 > \theta_2,\ \theta_1 \cdot \theta_2 > 0$ does not allow an orthogonal encoding. Therefore, we first formally define *linear constraints* as a class of constraints for which SICS will definitely work.

Definition 1 (Linear constraints)**.** *Let θ be an n-dimensional parameter vector, $\gamma_i \in \mathbb{R}^n$ an n-dimensional real vector, and Ω_i a non-empty convex subset of $\mathbb{R}$. Then we call a constraint with feasible set*

$$C_i = \{\theta : \gamma_i'\theta = \omega_i,\ \ \omega_i \in \Omega_i\}$$

a partial linear constraint $\mathcal{C}_i$ with criterion set Ω_i and coefficient vector γ_i. Further, let $\mathcal{C} = \{\mathcal{C}_j : 1 \leq j \leq k\}$ be a set of k partial linear constraints, let

$$\Gamma := \begin{pmatrix} \gamma_1' \\ \vdots \\ \gamma_k' \end{pmatrix}$$

be the $k \times n$ matrix of the concatenated and transposed coefficient vectors, and let

$$\Omega := \Omega_1 \times ... \times \Omega_k$$

be the Cartesian product of the criterion sets of the k partial linear constraints. Then we call the constraint $\mathcal{C}$ with feasible set

$$C = \{\theta : \Gamma\theta = \omega,\ \ \omega \in \Omega\}$$

a linear constraint with coefficient matrix Γ and criterion set Ω. Further, we define a linear equality constraint as a linear constraint whose criterion set contains only one element, $\Omega = \{a\}$, $a \in \mathbb{R}^k$; and we define a linear inequality constraint as a linear constraint

TABLE 6.1: Matrix representation of example constraints. $\theta \in \mathbb{R}^4$, $c \in \mathbb{R}$.

Constraint	$\Gamma\theta = \omega$	
$\theta_1 = 0$	$\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \theta = 0$	
$\theta_1 = \theta_2$	$\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \theta = 0$	
$\theta_1 > \theta_2 > \theta_3 > \theta_4$	$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \theta = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix}$,	$\omega_1, \omega_2, \omega_3 > 0$
$\theta_1 = \theta_2 > \theta_3 > \theta_4$	$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \theta = \begin{pmatrix} 0 \\ \omega_2 \\ \omega_3 \end{pmatrix}$,	$\omega_2, \omega_3 > 0$
$\frac{\theta_1+\theta_2+\theta_3+\theta_4}{4} = c$	$\begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \theta = c$	
$\theta_1 - \theta_2 > 0$, $\theta_3 - \theta_4 > \theta_1 - \theta_2$	$\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 1 & -1 \end{pmatrix} \theta = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$,	$\omega_1, \omega_2 > 0$
$\theta_2 - \theta_1 < \theta_3 - \theta_2 < \theta_4 - \theta_3$	$\begin{pmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{pmatrix} \theta = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$,	$\omega_1, \omega_2 > 0$
$\theta_1 - \theta_2 \in [-c, c]$	$\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \theta = \omega_1$,	$\omega_1 \in [-c, c]$

whose criterion set is the Cartesian product of open intervals on $\mathbb{R}$, $\Omega = (l_1, u_1), \times ... \times (l_k, u_k), -\infty \leq l_i < u_i \leq \infty, 1 \leq i \leq k$.

Linear constraints are those that can be described by equations and inequalities that only involve weighted sums and differences of parameters. This covers a majority of the constraints typically tested in research, in particular all equality and ordinal constraints; see Table 6.1 for examples.

The distinction between linear equality and inequality constraints is important, as equality constraints decrease the effective dimension of the parameter. This will become important below, when we derive the general solution to the reparameterization approach.

Note that a linear constraint is fully specified by its coefficient matrix Γ and the criterion set Ω. In practice, however, it is more common to specify constraints using symbolic notation, such as $\theta_1 < \theta_2 + 2\theta_3 = \theta_4$. In this case, Γ and Ω can be systematically derived via simple algebraic rules; see Observation 1 in Appendix 6.B for a short guide.

Next, we introduce the most important criterion for the application of SICS. We can use the method to estimate the marginal likelihood of a constrained hypothesis whenever this criterion is fulfilled.

Definition 2 (Linearly independent constraints)**.** *We call a linear constraint $\mathcal{C}$ consisting of k partial linear constraints linearly independent if the coefficient vectors $\gamma_j, 1 \leq j \leq k$ are linearly independent.*

Loosely speaking, a set of vectors is linearly independent if no vector in the set can be "built" from the others by scaling and adding them together. In practical terms, linearly independent constraints are those that allow us to form an invertible function τ. In more intuitive terms, linear independence ensures that the constraint contains no redundancies, contradictions, or partial constraints that are impossible to encode in an orthogonal way.

Lastly, we define how two linear constraints $\mathcal{A}$ and $\mathcal{B}$ can be combined into a composite constraint.

Definition 3 (Intersection of linear constraints)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be linear constraints on $\Theta \subset \mathbb{R}^n$ consisting of k_A and k_B partial linear constraints, with coefficient matrices $\Gamma_A \in \mathbb{R}^{k_A \times n}$ and $\Gamma_B \in \mathbb{R}^{k_B \times n}$, and criterion sets $\Omega_A \subset \mathbb{R}^{k_A}$ and $\Omega_B \subset \mathbb{R}^{k_B}$. Then the intersection $\mathcal{A} \cap \mathcal{B}$ is defined as the linear constraint with feasible set*

$$C_A \cap C_B = \left\{ \theta : \begin{pmatrix} \Gamma_A \\ \Gamma_B \end{pmatrix} \theta = \begin{pmatrix} \omega_A \\ \omega_B \end{pmatrix}, \ \omega_A \in \Omega_A, \omega_B \in \Omega_B \right\}.$$

For $k_B = 0$, we further define $\mathcal{A} \cap \mathcal{B} = \mathcal{A}$.

Thus, we simply combine two linear constraints by concatenating their coefficient matrices and criterion sets. This definition allows us to state all our results for pure equality constraints, pure inequality constraints, and mixed constraints.

Reparameterization to Ensure Constraint Satisfaction

We now state our two central results. First, we show that the desired reparameterization, or "inversion" of the roles of θ and ω, is possible whenever the constraint of interest is linearly independent. We then derive the prior and posterior on ω, as well as representations of the normalizing constants in terms of ω. This allows us to estimate the marginal likelihood by sampling from the constrained prior and posterior.

To formalize the reparameterization strategy introduced in the example, our first lemma guarantees that such a reparameterization exists and is invertible - as long as the constraint is linearly independent.

Lemma 1 (Reparameterization of constrained parameter)**.** *Let $\mathcal{A}$ be a linear equality constraint on the parameter vector $\theta \in \Theta \subset \mathbb{R}^n$ with $k_A \times n$ coefficient matrix Γ_A and criterion set $\Omega_A = \{a\}, a \in \mathbb{R}^{k_A}$. Also, let $\mathcal{B}$ be a linear inequality constraint with $k_B \times n$ coefficient matrix Γ_B and criterion set $\Omega_B \subset \mathbb{R}^{k_B}$. Additionally, let the intersection $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$ be linearly independent. Then we have the following:*

1. *There exists a $(n - k_A - k_B) \times n$ matrix U with rows that are linearly independent of the rows of $\begin{pmatrix} \Gamma_A \\ \Gamma_B \end{pmatrix}$.*
2. *There exists a bijective mapping τ between the feasible set $C = C_A \cap C_B$ and the transformed parameter space $\Omega = \Omega_B \times \mathbb{R}^{n-k_A-k_B}$ of the form*

$$\tau(\theta) = \begin{pmatrix} \Gamma_B \\ U \end{pmatrix} \theta, \quad \theta \in C_A \cap C_B$$

$$\tau^{-1}(\omega) = M \begin{pmatrix} a \\ \omega \end{pmatrix}, \quad \omega \in \Omega$$

where

$$M := \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix}^{-1}.$$

Proof: *See Appendix 6.B, Proof 1.*

The lemma ensures that we can find an invertible function τ, even in those cases where we have fewer constraints than parameters. To understand the significance of this, recall our motivating example: Here, we had a three-dimensional parameter θ, and the three partial constraints $\theta_1 > \theta_2$, $\theta_2 > \theta_3$, and $\theta_3 > 1$. Thus, the criterion vector ω was also three-dimensional, and we could uniquely express θ as a function of ω by inverting the matrix Γ. However, if we had only had two partial constraints, we would not have been able to do this, since it is impossible to express a three-dimensional parameter as a function of a two-dimensional parameter (trying to do so would be similar to trying to completely fill three-dimensional space with a two-dimensional plane). Technically, we would have been confronted with the impossibility of inverting the Γ matrix, since it has fewer rows than columns. However, we can easily solve this problem by extending Γ with a matrix U whose rows are linearly independent of the rows of Γ, thus creating an invertible square matrix. The parameter vector ω is extended by as many auxiliary parameters as U has rows. Intuitively, this can be understood as adding auxiliary "dummy" constraints that are always satisfied to the original constraint. As proved in our lemma, it is always possible to find such a matrix U. See Appendix 6.B, Observation 2 for an algorithm to compute a suitable matrix U.

Our second lemma establishes how to compute the prior, posterior, and marginal likelihood of the reparameterized model. This result ensures that we can use standard sampling and integration techniques - even when the constrained parameter space is lower-dimensional.

Lemma 2 (Prior, posterior, and normalizing constants after reparameterization)**.** *We use the same terms and definitions as in Lemma 1. In particular, recall that C is the feasible set of the constraint in question, and $\tau^{-1}(\omega) = M \begin{pmatrix} a \\ \omega \end{pmatrix}$. Additionally, let M_Ω be the $n \times (n - k_A)$ matrix of the last $n - k_A$ columns of M, let $p(D \mid \cdot)$ be the likelihood function defined on Θ, π be the prior on θ, and $\pi(\cdot \mid D)$ the posterior after observing data D. Then the prior and posterior on ω are given by*

$$\pi_\Omega(\omega) = \frac{\pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}}{\pi(\theta \in C)}$$

and

$$\pi_\Omega(\omega \mid D) = \frac{p(D \mid \tau^{-1}(\omega))\, \pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}}{p(D \mid \theta \in C)},$$

where $\mid \cdot \mid$ is the absolute matrix determinant. The normalizing constants are given by

$$\pi(\theta \in C) = \int_\Omega \pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}\, d\lambda_\Omega(\omega)$$

and

$$p(D \mid \theta \in C) = \frac{\pi(D \cap \theta \in C)}{\pi(\theta \in C)},$$

where

$$\pi(D \cap \theta \in C) = \int_{\Omega} p(D \mid \tau^{-1}(\omega))\, \pi(\tau^{-1}(\omega)) \sqrt{\mid M_{\Omega}' M_{\Omega} \mid}\, d\lambda_{\Omega}(\omega),$$

and where $d\lambda_{\Omega}(\omega)$ *is the Lebesgue measure on* $\Omega \subset \mathbb{R}^{n-k_A}$.

Proof: *See Appendix 6.B, Proof 2.*

If the constraint contains strict equalities, the feasible set is of a lower dimension than the original parameter space. Lemma 2 explains how in this case, the original prior or posterior is constrained to this lower dimensional space. A useful way to think about this is in geometric terms: the feasible set behaves like a "cut" through the original parameter space, and the induced prior can be thought of as the corresponding "slice" of the original prior. The normalization constants ensure that integrating the induced prior over the slice produces 1.

Together, Lemma 1 and Lemma 2 provide the theoretical foundation for SICS: We are guaranteed an invertible reparameterization that respects the constraint structure (Lemma 1), and we can express all relevant distributions and marginal likelihoods in terms of the transformed parameter (Lemma 2).

Marginal Likelihood Estimation

Given all the previous results, it is now straightforward to estimate the marginal likelihood using the following steps:

1. Using an MCMC algorithm such as Hamiltonian Monte Carlo, draw N_0 samples $\hat{\omega}_j, 1 \leq j \leq N_0$ from the reparameterized prior π_{Ω}, and N_1 samples $\check{\omega}_l, 1 \leq l \leq N_0$ from the reparameterized posterior $\pi_{\Omega}(\cdot \mid D)$.
2. Using an algorithm such as bridge sampling, estimate $K_1 \approx \pi(\theta \in C)$ from the prior samples $\hat{\omega}_j$, and $K_2 \approx \pi(D \cap \theta \in C)$ from the posterior samples $\check{\omega}_l$.
3. Estimate the marginal likelihood as $p(D \mid \theta \in C) \approx K_2 / K_1$.

Thus, once the model has been reparameterized, marginal likelihood estimation proceeds exactly as in any standard application of Bayesian computation. By making the constraint an integral part of the model, SICS eliminates the need to sample from the whole space and assess constraint satisfaction, as required by the encompassing prior methods.

6.3.4 Precision

The SICS estimator is derived as the ratio of two normalizing constants. Thus, its precision depends on the precision of the method used to estimate these constants. When bridge sampling is used, the precision of the estimates depends on the number of parameters, and the similarity between the constrained prior/posterior and the chosen proposal distribution (Gronau et al., 2017; Meng & Wong, 1996). Given the mathematical complexity of bridge sampling, it is not easy to predict when SICS will perform with higher or lower precision. However, the following observations make it a promising choice for estimating evidence for constrained hypotheses:

- **Bridge sampling is (close to) optimal:** As shown by Meng and Wong (1996), bridge sampling is asymptotically optimal among all importance sampling algorithms that use a particular proposal distribution g. This means that for very large MCMC samples, bridge sampling has a precision that is close to that of the best estimator that could possibly be obtained when using the proposal distribution g.
- **SICS can be expected to outperform the EP methods:** The unconditional EP method estimates the normalization constants by sampling from the unconstrained prior and posterior and calculating the proportion of samples that fall within C. This is equivalent to Monte Carlo integration without the use of a proposal distribution. However, unweighted sampling is known to produce sub-optimal precision. The variance can be greatly reduced by choosing an appropriate proposal distribution that better matches the target (Gronau et al., 2017). The problem is exacerbated by the fact that the unconstrained and constrained distributions by design have only partially overlapping support, a situation that is particularly detrimental to precision. The conditional EP method mitigates this issue to some extent by sequentially estimating the terms $\mathbb{P}(\theta \in C_i \mid \theta \in C_1, ..., \theta \in C_{i-1})$ and $\mathbb{P}(\theta \in C_i \mid \theta \in C_1, ..., \theta \in C_{i-1}, D)$ by sampling from the prior and posterior, respectively, conditioned on the constraints 1 to $i-1$. While this improves precision, it is still suboptimal due to the lack of a variance-reducing proposal distribution and the partial overlap of the supports. In contrast, SICS can be used with bridge sampling, which uses a proposal distribution to reduce variance in an asymptotically optimal way. The reparameterization ensures that the target and proposal distributions have fully overlapping support, further increasing precision.
- **Precision does not depend on the size of the feasible set:** Unlike EP methods, SICS is agnostic about the size of the feasible set C. This is due to the fact that all samples must fall in C due to the reparameterization.

Taken together, these features suggest that SICS should outperform the EP methods in the majority of cases. In the next section, we will put this claim to the test by analyzing an empirical dataset and comparing the performance of SICS with that of its predecessors.

6.4 Empirical Example: Decreasing Learning Rates in the Reinforcement Learning Model

To illustrate the use of SICS and compare it to existing methods, we consider the case of testing ordinal constraints on the learning rate parameters of a reinforcement learning (RL) model, applied to data from a two-armed bandit learning task. Originally developed in AI research (Sutton & Barto, 2018), RL models have become popular tools in behavioral and neural data analysis (Collins & Shenhav, 2021; Pessiglione et al., 2006; Schultz et al., 1997), with particular impact in computational psychiatry and social neuroscience (Adams et al., 2015; Lockwood & Klein-Flügge, 2020).

However, from the perspective of statistical estimation, RL models are quite complex, as they have a nonlinear structure and a likelihood function that must be evaluated algorithmically, as it cannot be reduced to a simple analytical expression. Moreover, there have been no developments towards testing constrained hypotheses in RL models. By demonstrating the ease of use and effectiveness of SICS in this case, we

hope to show that our method can be easily applied to all kinds of models, both simple and complex.

In a typical reinforcement learning task, participants repeatedly choose between two or more options, where each option has a certain probability to result in a positive outcome (e.g., reward, or avoidance of punishment) or a negative outcome (e.g., no reward, or punishment). The RL model most commonly used to analyze participants' choice sequences is a variant of the classical Rescorla-Wagner model (Rescorla & Wagner, 1972), which suggests that a participant's choice between options A and B at time t depends on the subjective values of the two options, v_t^A and v_t^B. In most applications, it is assumed that the subjective values are mapped to choice probabilities by the softmax function,

$$\mathbb{P}(\text{choice}_t = A \mid v_t^A, v_t^B) = \frac{1}{1 + \exp(-\beta(v_t^A - v_t^B))} \quad ,$$

where $\beta \in \mathbb{R}^+$ is the free inverse temperature parameter that determines the function's sensitivity to the value difference $v_t^A - v_t^B$. After each choice, the subjective value of the chosen option is updated according to the rule

$$v_{t+1}^{o(t)} = v_t^{o(t)} + \alpha(R_t - v_t^{o(t)}), \tag{6.13}$$

where $o(t)$ is the option chosen at trial t, R_t is the value of the outcome[4] received at trial t, and where $\alpha \in [0, 1]$ is the free *learning rate* parameter. The learning rate controls how much the last result influences the subjective values: high values lead to fast updating, while low values lead to a continuous integration of all previous results.

Several recent studies have shown that the learning rate and inverse temperature parameters of the RL model can differ between different populations, such as clinical vs. non-clinical subpopulations, (Montague et al., 2012; Vandendriessche et al., 2022) and between different contexts, such as self-oriented vs. prosocial learning (Lengersdorff et al., 2020; Lockwood et al., 2016). Within the same experimental condition, however, the learning rate α is typically assumed to remain constant over time. However, it is also possible that people update their subjective value of a choice option to a lesser extent the more information they have already collected about it. This would be reflected in learning rates that *decrease* with each exposure to a choice-outcome pairing. For example, assuming that people update their beliefs about an option's value using Bayes' rule, the learning rate would be inversely proportional to the number of outcomes received (see Appendix 6.C for a short proof). Modern Bayesian cognitive models of learning, such as the hierarchical Gaussian filter (C. Mathys et al., 2011), also predict decreasing learning rates in environments where contingency between choices and outcomes is not expected to change. Similarly, decreasing learning rates are a fundamental feature of RL models used in machine learning, often leading to better model fit and convergence than constant learning rates (Y. Chen et al., 2021; Mounjid & Lehalle, 2023).

[4] In settings with dichotomous outcomes, arbitrary values can be chosen for the two outcomes, as long as the value of the "positive" outcome is strictly larger than the value of the "negative" outcome. A typical choice is $R_t = 1$ for positive, and $R_t = 0$ for negative outcomes. A different choice of these values affects the model only through a rescaling of the inverse temperature parameter β (see supplementary material of Lengersdorff et al., 2020, accessible at https://osf.io/h9txe).

In the language of statistical modeling, decreasing learning rates can be represented by a k-dimensional parameter subject to the ordinal constraint

$$\alpha_1 > \alpha_2 > ... > \alpha_k,$$

where k is the number of distinct and decreasing learning rates, α_i is the learning rate applied to the i-th outcome received from a choice option, and α_k is the learning rate applied to all outcomes after and including the k-th outcome. In our exemplary analyses, we will test the hypotheses that human reinforcement learning is characterized by up to eight different and decreasing learning rates. We therefore formulate eight hypotheses of the form

$$\begin{aligned} \mathcal{H}_1 &: \alpha_1 \text{ unconstrained} \\ \mathcal{H}_2 &: \alpha_1 > \alpha_2 \\ &\vdots \\ \mathcal{H}_8 &: \alpha_1 > \alpha_2 > \alpha_3 > ... > \alpha_8, \end{aligned} \tag{6.14}$$

where $\alpha_i \in [0, 1]$ for $1 \leq i \leq 8$.

In the following, we will first demonstrate the use of SICS, and compare it to the unconditional and conditional encompassing prior approaches, by analyzing an exemplary learning task dataset from a single participant. We will then illustrate the generalizability of SICS by testing our hypotheses in the multilevel framework, analyzing a dataset from a larger number of participants. Of central interest to our analyses is whether or not one or more of the models allowing for decreasing learning rates ($\mathcal{H}_2$ to $\mathcal{H}_8$) outperform the conventional RL model with only a single learning rate ($\mathcal{H}_1$).

6.4.1 Comparison of Methods in Single-Subject Analysis

Method

We analyzed the two-armed bandit task dataset of Schaaf et al., 2023, which is openly available at https://osf.io/pe23t/. Participants completed a standard reinforcement learning paradigm, in which eight pairs of choice options were presented. Each choice option had a fixed probability of leading to a positive monetary outcome, and these probabilities were complementary within pairs (70% vs. 30% in half of the pairs; 80% vs. 20% in the other half). See Schaaf et al. (2023) for more details. To compare the methods, we analyzed the data from participant 1. An illustration of the participant's choice and outcome sequence is shown in Figure 6.3.

To compare SICS with the unconditional and conditional EP methods, we repeatedly estimated the marginal likelihoods of the eight hypotheses described in (6.14) using all three methods. To estimate marginal likelihoods using the EP methods, we used the following relation:

$$\text{BF}_{C\Theta} = \frac{p(D \mid \theta \in C)}{p(D)} \implies p(D \mid \theta \in C) = \text{BF}_{C\Theta} \cdot p(D).$$

Thus, the marginal likelihood $p(D \mid \theta \in C)$ can be obtained from $\text{BF}_{C\Theta}$ by multiplying it by the marginal likelihood of the encompassing model $p(D)$. Here, we achieved

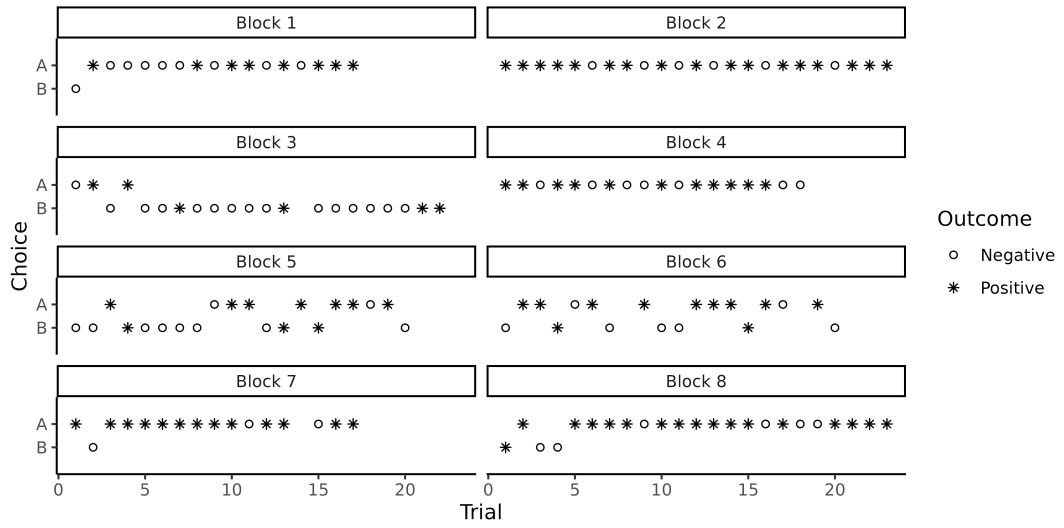

FIGURE 6.3: Two-armed bandit task data of the example participant. Missing markers at a trial indicate a missing response.

this by estimating $\text{BF}_{C\Theta}$ using the respective EP method, and $p(D)$ using bridge sampling.

We formulated a uniform prior on each of the learning rates α_i, and an exponential prior on the inverse temperature parameter β. We repeated the estimation of each quantity 100 times. The prior and posterior samples required for marginal likelihood estimation were drawn using custom models written in STAN (Stan Development Team, 2024), using R as interface. All STAN and R code is accessible on https://github.com/LukasLengersdorff/sics-analyses. To estimate the marginal likelihood $p(D \mid \theta \in C)$ via the EP methods, we estimated the terms $\pi(\theta \in C \mid D)$ and $\pi(\theta \in C)$ as described in (6.6), as well as the marginal likelihood of the encompassing $p(D)$. For estimation via SICS, we estimated the terms $\pi(\theta \in C)$ and $\pi(D \cap \theta \in C)$, as described in Lemma 2. To make the results comparable across estimation methods, the estimation of each of these terms was based on a total of $15,000$ samples. Since the conditional EP method estimates k partial BFs for k partial inequality constraints (see (6.8)), the sample size used to estimate the nominator and denominator of each partial BF was set to $\lceil 15000/k \rceil$, where the ceiling function $\lceil \cdot \rceil$ rounds up to the next integer. Note that hypothesis $\mathcal{H}_1$ has only a single parameter without any constraints. Therefore, its marginal likelihood was estimated by simple bridge sampling.

As a measure of computational resources used, we recorded the CPU time (i.e. the time the computer actively performed computations) of each estimation, using the R function `proc.time`. As a measure of relative efficiency, we calculated the ratio of precision to mean CPU time per method,

$$\text{eff}_{\text{method}} := \frac{SD_{\text{method}}^{-1}}{\bar{t}_{\text{method}}},$$

where SD_{method} is the standard deviation of the method's marginal likelihood estimates, and $\bar{t}_{\text{method}}$ is the mean CPU time consumed by the method.

Results

The results are shown in Figure 6.4. On average, all three methods produced similar marginal likelihood estimates, with the exception of the unconditional EP for $\mathcal{H}_8$, which failed to produce an estimate with the given sample size. However, for all hypotheses, SICS had substantially lower sampling variability, and this precision advantage became more significant as more constraints were tested. See Table 6.D.1 in Appendix 6.D for numerical details.

Across all hypotheses, SICS also showed the highest computational efficiency, with particularly large gains for highly constrained models. SICS was up to 25 times more efficient than the conditional EP, and managed to produce accurate estimates in cases where the unconditional EP failed. See 6.D.2 in Appendix 6.D for complete efficiency indices.

To summarize the comparison of methods, we found that SICS is an effective method for estimating the marginal likelihood, outperforming the EP methods in terms of precision and efficiency. These results suggest that constrained hypothesis testing can greatly benefit from the two main features of SICS: re-parameterization to ensure that the tested constraint is satisfied, and variance reduction through the use of efficient algorithms for estimating marginal likelihoods, such as bridge sampling.

6.4.2 SICS in Multilevel Analysis

Method

To demonstrate the application of SICS in the multilevel framework, we analyzed the first 20 participants of the data set of Schaaf et al. (2023). For each hypothesis $\mathcal{H}_k$ (k different and decreasing learning rates), we assumed that the behavior of participant j was characterized by a k-dimensional participant-level vector of learning rates $\alpha_j = (\alpha_{j1}, ..., \alpha_{jk})$, and an inverse temperature parameter β_j. At the population level, we assumed that the logits of α_j and the logarithm of β_j follow a multivariate normal distribution

$$\begin{pmatrix} \text{logit}\, \alpha_j \\ \log \beta_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \Sigma \right),$$

where $\mu_\alpha = (\mu_{\alpha_1}, ..., \mu_{\alpha_k})$ and μ_β are the population-level mean parameters, and Σ is the $(k+1)$-dimensional variance-covariance matrix. The latter has been further parameterized as

$$\Sigma = \text{diag}(\sigma_{\alpha_1}, \sigma_{\alpha_2}, ..., \sigma_\beta) \cdot R \cdot \text{diag}(\sigma_{\alpha_1}, \sigma_{\alpha_2}, ..., \sigma_\beta),$$

where the elements of the diagonal matrix are population-level standard deviations, and R is a $(k+1)$ dimensional correlation matrix. To complete our model specification, we have formulated the following priors on the population parameters:

$$\begin{aligned} \mu_{\alpha_j}, \mu_\beta &\sim \mathcal{N}(0, 1/2) \\ \sigma_{\alpha_j}, \sigma_\beta &\sim \text{Lognormal}(0, 1/2) \\ R &\sim \text{LKJ}(1), \end{aligned}$$

where LKJ is the LKJ prior on correlation matrices (Lewandowski et al., 2009). We considered two different ways to implement constrained hypothesis testing in the multilevel framework. First, we tested a family of mean-constrained multilevel

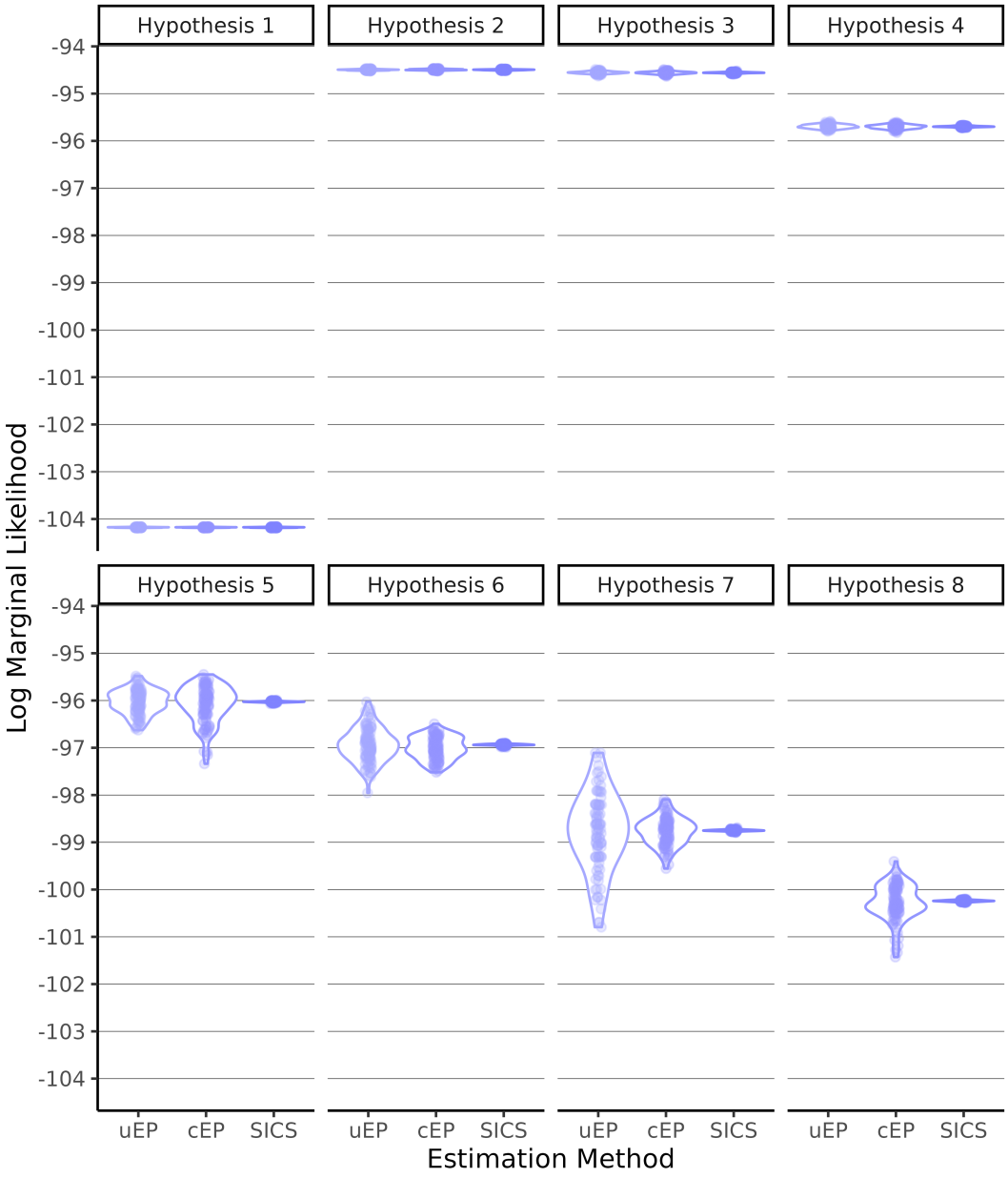

FIGURE 6.4: Estimation of the log marginal likelihood of Hypotheses $\mathcal{H}_1$ to $\mathcal{H}_8$ using the unconditional encompassing prior method (uEM), conditional encompassing prior method (cEM), and Sampling-in-Constrained-Space (SICS). Estimation was repeated 100 times per method and hypothesis. **Notes:** For Hypothesis 1, the estimation procedure was the same for all three methods, as no constraints needed to be evaluated. For Hypothesis 8, all estimates using uEM were invalid ($+\infty$, $-\infty$, or $\infty - \infty$).

models ($ML_{\text{mean con}}$) in which only the population-level mean parameters were constrained:

$$\mathcal{H}_k \text{ under } ML_{\text{mean con}} : \mu_{\alpha_1} > ... > \mu_{\alpha_k}.$$

Thus, in these models, only the average of the learning rates, but not each participant-level parameter vector, was required to satisfy the constraint. Second, we evaluated a family of fully constrained multilevel models ($ML_{\text{full con}}$) in which the learning rates of each individual participant followed the constraint:

$$\mathcal{H}_k \text{ under } ML_{\text{full con}} : \alpha_{j1} > ... > \alpha_{jk}, \text{ for all } 1 \leq j \leq 20.$$

These two model families represent different theories: $ML_{\text{full con}}$ proposes that decreasing learning rates are a universal feature of human reinforcement learning, whereas $ML_{\text{mean con}}$ suggests that this pattern is merely a general tendency that allows for significant inter-individual variation (c.f. Haaf & Rouder, 2017).

In addition, we tested a family of unconstrained multilevel models (ML_{free}) in which learning was characterized by k time-dependent, but not necessarily decreasing, learning rates. We included these as a comparative benchmark: If we find the best model fit for some $\mathcal{H}_k$ under ML_{free}, rather than under $ML_{\text{mean con}}$ or $ML_{\text{full con}}$, this would indicate that it is important to consider k different learning rates, but not that they are also decreasing.

For each family of models and hypotheses, the marginal likelihood was estimated via SICS, using $40,000$ samples from the constrained prior (4 chains with $10,000$ samples per chain), and $8,000$ samples from the constrained posterior (4 chains with $2,000$ samples per chain).

Results

The results are shown in Figure 6.5, panel A. Detailed numerical results can be found in Table 6.D.3. Perhaps most strikingly, all three model families show extremely strong evidence for the hypothesis that learning is characterized by more than one learning rate. The log BF for $\mathcal{H}_2$ compared to $\mathcal{H}_1$ (i.e., the difference in the log marginal likelihoods) is between 93 and 95 for all three model families, indicating that the observed data are more than $\exp(93) = 2.45 \cdot 10^{40}$ times more likely to be modeled by a second learning rate than by a single learning rate. The highest log marginal likelihood of -1414.2 was found for $\mathcal{H}_6$ under $ML_{\text{mean con}}$ (the means of learning rates are decreasing). Comparing this to $\mathcal{H}_1$ gives a BF of 208.2, indicating that the data are $\exp(208.2) = 2.63 \cdot 10^{90}$ times more likely under a model with six decreasing mean learning rates than under a single learning rate. These results suggest very strongly that reinforcement learning models with a single time-invariant learning rate provide suboptimal fit to human behavior - at least in the present sample.

Furthermore, for all hypotheses with more than one learning rate, $ML_{\text{mean con}}$ outperformed ML_{free}, indicating that it is not only important to model different learning rates, but also that these are decreasing in mean. See Figure 6.5, panel B. For example, the BF of the overall best-fitting hypothesis $\mathcal{H}_6$ under $ML_{\text{mean con}}$ against $\mathcal{H}_6$ under ML_{free} is $\exp(5.92) = 372.4$, suggesting that the data are about 372 times more likely under a model in which we assume that the means of the six learning rates are decreasing rather than unordered.

We can also observe that there was generally much less evidence for $ML_{\text{full con}}$, compared to the other models. This means that models in which the learning rates of

each individual participant were constrained to be decreasing performed worse than models in which the learning rates were allowed to vary more freely. Thus, the data suggest that strictly decreasing learning rates are not a universal feature of human learning, and that there is substantial individual deviation from this pattern.

In summary, our analyses provide strong evidence that human reinforcement learning is characterized by decreasing learning rates. However, this result is only true on average - individuals may exhibit learning rates that deviate from this strictly decreasing pattern. Most importantly for the present paper, our empirical analyses demonstrate that SICS can be used effectively to test constraints on different levels of the hierarchical structure in a multilevel framework.

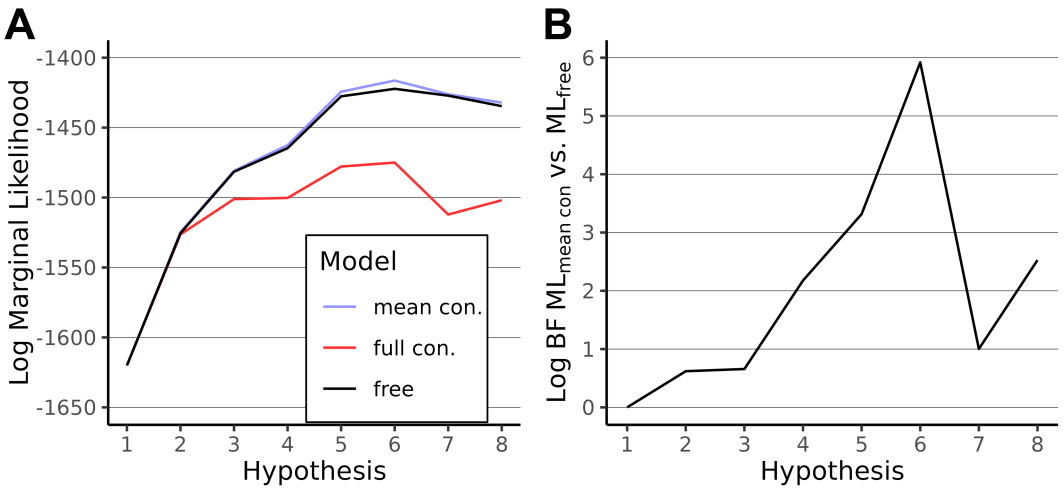

FIGURE 6.5: R
esults of constrained multilevel RL model analysis. **A)** Estimates of the log marginal likelihood of Hypotheses $\mathcal{H}_1$ to $\mathcal{H}_8$ in constrained multilevel models. **B)** Log Bayes factors of mean-constrained multilevel models vs. unconstrained models.

6.5 Outlook and Discussion

With SICS, we have introduced an efficient and widely generalizable method for estimating evidence for ordinal constraints in statistical models amenable to standard numerical integration techniques. The main component of the method is to reparameterize the statistical model so that the sampled parameters always satisfy the ordinal constraints. Importantly, this step is independent of the remaining structure of the model and its priors, and thus SICS can be implemented for all types of statistical models and specified priors.

To make SICS available to a wide range of researchers, an easy-to-use software implementation is needed. In our view, such software should do the following

1. Provide a representation of linear constraints. Users should be able to flexibly specify and combine the constraints they want to test.
2. Automatically calculate and specify all necessary data. This includes deriving the matrix representation of a constraint from its formulaic description.
3. Automatic or semi-automatic adaptation of existing model implementations (e.g., MCMC samplers) to sample from the transformed rather than the original parameter space.

The code we developed for the analyses presented in this paper fulfills points (1) and (2), and we aim to publish it as a standalone R package in the near future. Regarding point (3), automatic conversion of existing models into SICS-compliant versions does not yet exist. However, we provide clear templates on how to adapt existing STAN model specifications to enable SICS. In the future, we see two ways in which automatic model conversion could be achieved. First, SICS could be implemented as a plug-in to software such as `brms` (Bürkner, 2017), which automatically writes and compiles STAN model code based on user specifications. Such a plug-in would simply need to redefine the original model parameters as a function of the new parameters in the generated code. Second, SICS could take advantage of the modularity of probabilistic programming languages such as NumPyro (Phan et al., 2019) and Edward2 (Tran et al., 2018), which allow flexible re-parameterization of existing models (Gorinova et al., 2020). As an alternative, SICS-compliant models could also be collected in an online repository, allowing researchers to share their custom models with others.

In summary, SICS is a novel and highly general approach to Bayesian hypothesis testing under constrained models. Through empirical applications, we demonstrated that SICS addresses key limitations of existing methods, including computational inefficiency and scalability issues, while ensuring precision in marginal likelihood estimation. By applying SICS to reinforcement learning models with constrained learning rates, we validated its versatility and robustness in complex statistical frameworks. These results highlight SICS as a valuable addition to the inferential toolkit.

Data and Code availability

The dataset of Schaaf et al., 2023, analyzed in the empirical example, is openly accessible at https://osf.io/pe23t/. Code to reproduce all reported analyses is available on Github: https://github.com/LukasLengersdorff/sics-analyses.

Funding

This work was supported by a OeAD Marietta-Blau grant (MPC-2021-00158) awarded to LLL. MM was supported by the European Union (ERC, BAYESIAN P-NETS, #101040876). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgments

We would like to thank Eric-Jan Wagenmakers for valuable comments during the initial stage of this project.

6.6 Appendix

6.A Relative Error of the Encompassing Prior method

Unconditional Encompassing Prior Method

Consider the relative error of a probability estimator $\hat{P}$, defined as

$$R_{\hat{p}} = \frac{SD(\hat{P})}{P}$$

where $SD(\hat{P})$ is the estimator's standard deviation, and P is the probability of interest (Meng & Wong, 1996). In the unconditional EP method, the probabilities $\pi(\theta \in C)$ and $\pi(\theta \in C \mid D)$ are estimated through Monte Carlo integration using N samples, see 6.7. This amounts to estimating the respective probability P as the mean of N i.i.d. Bernoulli-distributed random variables with success probability P. The standard deviation of this estimate is $\sqrt{\frac{P(1-P)}{N}}$, and thus the relative error is

$$\begin{aligned} R_{\hat{p}} &= \sqrt{\frac{P(1-P)}{N}} \cdot \frac{1}{P} \\ &= \sqrt{\frac{1-P}{NP}}. \end{aligned}$$

For P close to zero, the relative error is approximately equal to the inverse square root of NP. By straightforward transformations, we may also calculate the sample size N needed to obtain a relative error below some threshold c:

$$N \geq \frac{1-P}{c^2 P}.$$

Conditional Encompassing Prior Method

Let C be the feasible set of a constrained hypothesis consisting of k inequality constraints C_i. We consider the estimation of the probability $\mathbb{P}(\theta \in C)$; equivalent results on the estimation of $\mathbb{P}(\theta \in C \mid D)$ can be derived anagolously.

Define $P := \mathbb{P}(\theta \in C)$ and $P_i := \mathbb{P}(\theta \in C_i \mid \theta \in \bigcap_{j=1}^{i-1} C_j)$. Then $P = \prod_{i=1}^{k} P_i$. In the conditional EP method, each P_i is estimated with the estimator

$$\hat{P}_i = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{\theta}_i),$$

where n is the size of the MCMC sample, and $\hat{\theta}_i$ are drawn from the respective prior conditioned on the constraints 1 to $i-1$.

As $\hat{P}_i$ is the mean of a binomially distributed variable, we have $\mathbb{E}[\hat{P}_i] = P_i$ and $\mathrm{Var}[\hat{P}_i] = \frac{P_i(1-P_i)}{n}$. The unconditional EP estimator of P is given by $\hat{P} = \prod_{i=1}^{k} \hat{P}_i$, and thus, its variance is the variance of the product of $\hat{P}_i$. By a well-known result

regarding the variance of the product of independent variables, we have

$$\begin{aligned}\mathrm{Var}(\hat{P}) &= \prod_{i=1}^{k}(\mathrm{Var}[\hat{P}_i] + \mathbb{E}[\hat{P}_i]^2) - \prod_{i=1}^{k}\mathbb{E}[\hat{P}_i]^2 \\ &= \prod_{i=1}^{k}\left(\frac{P_i(1-P_i)}{n} - P_i^2\right) - P^2 \\ &= \prod_{i=1}^{k}\left(\frac{P_i}{n}(1-P_i+nP_i)\right) - P^2.\end{aligned}$$

The relative error of the estimator is $R_{\hat{P}} = \dfrac{\sqrt{\mathrm{Var}(\hat{P})}}{P}$, but for ease of notation, we will for now consider its square,

$$R_{\hat{P}}^2 = \frac{\mathrm{Var}(\hat{P})}{P^2} = \frac{1}{P^2}\left(\prod_{i=1}^{k}\left(\frac{P_i}{n}(1-P_i+nP_i)\right) - P^2\right).$$

In the following, we use the fact that $P^2 = \prod_{i=1}^{k} P_i^2$, and that thus, P^2 can be split into its factors and distributed across the product:

$$\begin{aligned}R_{\hat{P}}^2 &= \frac{1}{P^2}\left(\prod_{i=1}^{k}\left(\frac{P_i}{n}(1-P_i+nP_i)\right) - P^2\right) \\ &= \prod_{i=1}^{k}\frac{1}{P_i^2}\left(\frac{P_i}{n}(1-P_i+nP_i)\right) - 1 \\ &= \prod_{i=1}^{k}\left(\frac{1-P_i+nP_i}{nP_i}\right) - 1 \\ &= \prod_{i=1}^{k}\left(\frac{1}{nP_i} + \frac{n-1}{n}\right) - 1 \\ &= \prod_{i=1}^{k}\left(\frac{1}{nP_i} + 1 - \frac{1}{n}\right) - 1.\end{aligned}$$

Due to the product, it is difficult to further simplify this expression. However, we can find a simpler expression for an upper bound of the squared relative error. First, because $\frac{1}{n}$ is positive, we have

$$R_{\hat{P}}^2 = \prod_{i=1}^{k}\left(\frac{1}{nP_i} + 1 - \frac{1}{n}\right) - 1 \leq \prod_{i=1}^{k}\left(\frac{1}{nP_i} + 1\right) - 1.$$

Next, define $P_{\min} := \min_{1\leq i\leq k} P_i$. Then

$$\prod_{i=1}^{k}\left(\frac{1}{nP_i} + 1\right) - 1 \leq \prod_{i=1}^{k}\left(\frac{1}{nP_{\min}} + 1\right) - 1 = \left(\frac{1}{nP_{\min}} + 1\right)^k - 1,$$

and thus

$$R_{\hat{P}}^2 \leq \left(\frac{1}{nP_{\min}} + 1\right)^k - 1.$$

To achieve an (unsquared) relative error that is below some threshold c, n must satisfy

$$c^2 \leq \left(\frac{1}{nP_{\min}} + 1\right)^k - 1.$$

By straightforward transformations, this condition is satisfied if

$$n \geq \frac{1}{P_{\min}} \frac{1}{\sqrt[k]{c^2+1} - 1}.$$

The right-hand term is an expression that is bounded above by a linear function in k. To see this, we use the Taylor expansion of the exponential function:

$$\begin{aligned}
\frac{1}{\sqrt[k]{c^2+1} - 1} &= \frac{1}{\exp(\log(c^2+1)/k) - 1} \\
&= \frac{1}{\sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{\log(c^2+1)}{k}\right)^m - 1} \\
&= \frac{1}{\sum_{m=1}^{\infty} \frac{1}{m!} \left(\frac{\log(c^2+1)}{k}\right)^m} \\
&\leq \frac{1}{\left(\frac{\log(c^2+1)}{k}\right)} \\
&= \frac{k}{\log(c^2+1)}.
\end{aligned}$$

Putting everything together, the relative error is certain to be smaller than c if

$$n \geq \frac{1}{P_{\min}} \cdot \frac{k}{\log(c^2+1)}.$$

Comparison of the two Encompassing Prior Methods

As stated above, the necessary sample size for the unconditional EP method N_u to achieve a relative error smaller than c must satisfy

$$N_u \geq \frac{1-P}{c^2 P}.$$

As $0 < P$ and $P_{\min}^k < P$, we have

$$\frac{1-P}{c^2 P} \geq \frac{1}{c^2 P_{\min}^k},$$

and thus

$$N_u \geq \frac{1}{c^2 P_{\min}^k}.$$

To compare this to the necessary sample size of the conditional EP, we should consider its total sample size (remember that n is the sample size used to estimate one P_i). When each P_i is estimated with the same number of samples n, we have $N_c := kn$

and thus $n = N_c/k$. Putting this together with the results from above,

$$\frac{N_c}{k} \geq \frac{1}{P_{\min}} \cdot \frac{k}{\log(c^2+1)}$$

and thus

$$N_c \geq \frac{k^2}{P_{\min} \log(c^2+1)}.$$

Thus, N_u is bounded below by a function that grows exponentially with regards to k, while the minimum N_c is proportional to only a second-degree polynomial in k. This explains why, for large k, the conditional EP can achieve much larger precision than the unconditional EP.

6.B Proofs

Observation 1. Given a linear constraint in symbolic representation, the coefficient matrix Γ and the criterion set Ω can be derived by following these steps:

1. Split the symbolic representation into as many partial constraints as there are relational symbols. E.g., $\theta_1 < \theta_2 + 2\theta_3 = \theta_4$ becomes $\theta_1 < \theta_2 + 2\theta_3, \theta_2 + 2\theta_3 = \theta_4$.
2. Collect all parameter components on the left side of each relation. E.g. $\theta_1 - \theta_2 - 2\theta_3 < 0, \theta_2 + 2\theta_3, \theta_2 + 2\theta_3 - \theta_4 = 0$.
3. Collect the coefficients of each parameter component in row vectors Γ_i. E.g. $\Gamma_1 = \begin{pmatrix} 1 & -1 & -2 & 0 \end{pmatrix}, \Gamma_2 = \begin{pmatrix} 0 & 1 & 2 & -1 \end{pmatrix}$.
4. Concatenate the row vectors to form the matrix Γ. E.g., $\Gamma = \begin{pmatrix} 1 & -1 & -2 & 0 \\ 0 & 1 & 2 & -1 \end{pmatrix}$
5. Replace the left side of the i-th relation with ω_i, and specify Ω_i as the set of numbers fulfilling this relation. E.g., $\omega_1 < 0 \implies \Omega_1 = (-\infty, 0)$ and $\omega_2 = 0 \implies \Omega_2 = \{0\}$.

Proof 1 (Lemma 1: Reparameterization of constrained parameter)**.**

ad (1): As $\mathcal{C}$ is linearly independent, the rows of the concatenated matrix $\Gamma := \begin{pmatrix} \Gamma_A \\ \Gamma_B \end{pmatrix}$ are independent. Thus, the rows of Γ, $(\Gamma_i')_{1 \leq i \leq (k_A + k_B)}$ form a basis for an $(k_A + k_B)$-dimensional subspace of $\mathbb{R}^n$. By the basis extension theorem, we can find vectors $(u_j)_{1 \leq j \leq (n - k_A - k_B)}$ that are linearly independent from the rows of Γ to extend $(\Gamma_i')_{1 \leq i \leq (k_A + k_B)}$ to a basis of $\mathbb{R}^n$. Setting $U := \begin{pmatrix} u_1' \\ \vdots \\ u_{(n-k_A-k_B)}' \end{pmatrix}$ gives the desired matrix.

ad (2): By definition, we have for any $\theta \in C_A \cap C_B$:

$$
\begin{aligned}
\theta \in (C_A \cap C_B) &\iff \begin{pmatrix} \Gamma_A \\ \Gamma_B \end{pmatrix} \theta = \begin{pmatrix} a \\ \omega_B \end{pmatrix}, \quad \omega_B \in \Omega_B \\
&\iff \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix} \theta = \begin{pmatrix} a \\ \omega_B \\ v \end{pmatrix}, \quad v \in \mathbb{R}^{n - k_A - k_B} \\
&\iff \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix} \theta = \begin{pmatrix} a \\ \omega \end{pmatrix}, \quad \omega \in \Omega \\
&\iff \theta = \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix}^{-1} \begin{pmatrix} a \\ \omega \end{pmatrix} = M \begin{pmatrix} a \\ \omega \end{pmatrix},
\end{aligned}
$$

where $M^{-1} = \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix}$ is invertible by the previous result. Because M is invertible, $\tau^{-1}(\omega) = M \begin{pmatrix} a \\ \omega \end{pmatrix}$ is a bijective map from Ω to $C_A \cap C_B$. Its inverse can be found by solving for ω:

$$\begin{aligned}
& \theta = M \begin{pmatrix} a \\ \omega \end{pmatrix} \\
\implies & M^{-1}\theta = \begin{pmatrix} a \\ \omega \end{pmatrix} \\
\implies & \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix} \theta = \begin{pmatrix} a \\ \omega \end{pmatrix} \\
\implies & \begin{pmatrix} 0_{(n-k_A)\times k_A} & I_{n-k_A} \end{pmatrix} \begin{pmatrix} \Gamma_A \\ \Gamma_B \\ U \end{pmatrix} \theta = \begin{pmatrix} 0_{(n-k_A)\times k_A} & I_{n-k_A} \end{pmatrix} \begin{pmatrix} a \\ \omega \end{pmatrix} \\
\implies & \begin{pmatrix} \Gamma_B \\ U \end{pmatrix} \theta = \omega.
\end{aligned}$$

Here, $0_{(n-k_A)\times k_A}$ is a $(n-k_A) \times k_A$ matrix filled with zeros, and I_{n-k_A} is the $(n-k_A)$-dimensional identity matrix. □

Observation 2 (Algorithm for computing linearly independent matrix U)**.** Let $\mathcal{C}$ be a linearly independent constraint with coefficient matrix $\Gamma \in \mathbb{R}^{k\times n}$. Then we can find a matrix $U \in \mathbb{R}^{(n-k)\times n}$ with rows that are linearly independent of the rows of Γ by the following algorithm:

1. Compute the orthogonal projection matrix on the complement of the row space of Γ, $P = I_n - \Gamma'(\Gamma\Gamma')^{-1}\Gamma$.
2. Compute the reduced row echelon form E of P.
3. Remove the last k rows of E, and set U to the remaining matrix.

Proof 2 (Lemma 2: Prior, posterior, and normalizing constants after reparameterization)**.** For $k_A = 0$ (no strict equality constraints), the results follows from a straightforward application of the change-of-variables formula for probability densities. For the general case ($k_A > 0$), we need to transfer the probability measure in Θ to the lower-dimensional spaces C and Ω. This is possible by results from differential geometry, see e.g. (Pennec, 2006). Very shortly, if C is a $(n-k_A)$ dimensional manifold in Θ parameterized by some function $\phi : \Omega \to \Theta$, then the probability measure on Ω implied by the probability measure π on Θ is given by

$$\pi_\Omega(\omega) = \pi(\phi(\omega))\sqrt{|\ J'J\ |}, \tag{6.15}$$

where J is the Jacobian of ϕ. In our case, the feasible set C is parameterized by the function τ^{-1}, which can be understood as the composition $h_2 \circ h_1$. Here, h_1 is given by

$$\begin{aligned} h_1 : \Omega &\to \{a\} \times \Omega \\ h_1(\omega) &= \begin{pmatrix} a \\ \omega \end{pmatrix} \end{aligned}$$

with Jacobian

$$J_{h_1} = \begin{pmatrix} 0_{k_A \times (n-k_A)} \\ I_{n-k_A} \end{pmatrix},$$

where $0_{k_A \times (n-k_A)}$ is a $k_A \times (n - k_A)$ matrix filled with zeros, and I_{n-k_A} is the $n - k_A$-dimensional identity matrix. Further, h_2 is given by

$$\begin{aligned} h_2 :& \{a\} \times \Omega \to C \\ h_2(v) =& Mv \end{aligned}$$

with Jacobian

$$J_{h_2} = M.$$

Thus, the Jacobian of the full transformation is

$$J = J_{h_2} \cdot J_{h_1} = M \begin{pmatrix} 0_{k_A \times (n-k_A)} \\ I_{n-k_A} \end{pmatrix} = M_\Omega.$$

The remaining results follow immediately from applying (6.15). For the transformed prior and posterior, we have

$$\begin{aligned} \pi_\Omega(\omega) &= \pi(\tau^{-1}(\omega) \mid \theta \in C) \sqrt{\mid M_\Omega' M_\Omega \mid} \\ &= \frac{\pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}}{\pi(\theta \in C)} \end{aligned}$$

and

$$\begin{aligned} \pi_\Omega(\omega \mid D) &= \pi(\tau^{-1}(\omega) \mid D, \theta \in C) \sqrt{\mid M_\Omega' M_\Omega \mid} \\ &= \frac{p(D \mid \tau^{-1}(\omega))\, \pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}}{p(D \mid \theta \in C)}. \end{aligned}$$

Similarly, the normalizing constants are given by

$$\begin{aligned} \pi(\theta \in C) &= \int_C \pi(\theta)\, \mathrm{d}\lambda_C(\theta) \\ &= \int_\Omega \pi(\tau^{-1}(\omega)) \sqrt{\mid M_\Omega' M_\Omega \mid}\, \mathrm{d}\lambda_\Omega(\omega). \end{aligned}$$

and

$$\begin{aligned} p(D \mid \theta \in C) &= \int_C p(D \mid \theta)\, \pi(\theta \mid \theta \in C)\, \mathrm{d}\lambda_C(\theta) \\ &= \int_\Omega p(D \mid \tau^{-1}(\omega)) \frac{\pi(\tau^{-1}(\omega))\sqrt{\mid M_\Omega' M_\Omega \mid}}{\pi(\theta \in C)}\, \mathrm{d}\lambda_\Omega(\omega) \\ &= \frac{\int_\Omega p(D \mid \tau^{-1}(\omega))\, \pi(\tau^{-1}(\omega))\sqrt{\mid M_\Omega' M_\Omega \mid}\, \mathrm{d}\lambda_\Omega(\omega)}{\pi(\theta \in C)} \\ &= \frac{\pi(D \cap \theta \in C)}{\pi(\theta \in C)} \end{aligned}$$

with $\pi(D \cap \theta \in C)$ as defined above. □

6.C Proof of Decreasing Learning Rates in Bayesian Learning

We consider i.i.d variables R_k, representing the outcome after choosing some choice option in an RL task for the k-th time. The variables R_k follow a Bernoulli distribution with unknown probability of success $\theta \in [0,1]$. Let θ have a beta prior with hyperparameters $a > 0, b > 0$, and let v_k be the posterior mean of p after observing realizations $R_1, ..., R_{k-1}$, $v_k = \mathbb{E}[\theta \mid R_1, ..., R_{k-1}]$. (Note that the k-th observation is left out, to conform with the notation from the RL model). Then by known characteristics of the beta distribution as conjugate prior of the Bernoulli distribution, we have

$$v_k = \frac{a + \sum_{i=1}^{k-1} R_i}{a + b + k - 1}.$$

For ease of notation, we define $S_k := a + \sum_{i=1}^{k-1} R_i$ and $K := a + b + k - 1$. Thus $v_k = \frac{S_k}{K}$. We are interested in finding α_k such that

$$v_{k+1} = v_k + \alpha_k(R_k - v_k).$$

Note that we may leave open the possibility that α_k may depend on the value of R_k. In the language of reinforcement learning models, α_k is the learning rate applied after the k-th outcome. Solving for α_k gives

$$\begin{aligned}
\alpha_k &= \frac{v_{k+1} - v_k}{R_k - v_k} \\
&= \frac{\frac{S_{k+1}}{K+1} - \frac{S_k}{K}}{R_k - \frac{S_k}{K}} \\
&= \frac{KS_{k+1} - (K+1)S_k}{K(K+1)R_k - (K+1)S_k} \\
&= \frac{K(S_k + R_k) - (K+1)S_k}{K(K+1)R_k - (K+1)S_k} \\
&= \frac{KR_k - S_k}{K(K+1)R_k - (K+1)S_k}.
\end{aligned}$$

For $R_k = 0$,

$$\begin{aligned}
\alpha_k &= \frac{0 - S_k}{0 - (K+1)S_k} \\
&= \frac{1}{K+1}.
\end{aligned}$$

For $R_k = 1$,

$$\begin{aligned}
\alpha_k &= \frac{K - S_k}{K(K+1) - (K+1)S_k} \\
&= \frac{(K - S_k)}{(K+1)(K - S_k)} \\
&= \frac{1}{K+1}.
\end{aligned}$$

Thus, $\alpha_k = \frac{1}{a+b+k}$, and is thus a strictly decreasing function in k.

6.D Additional Tables

TABLE 6.D.1: Results of the method comparison. Rows: Hypotheses. Columns: summary statistics per method. M: Mean of log marginal likelihood. SD: Standard deviation of log marginal likelihood. $\bar{t}$: Mean CPU time, in seconds. uEP: unconditional encompassing prior method. cEP: conditional encompassing prior method. SICS: Sampling in constrained space. -: no estimation possible, because no valid estimates were produced. For $\mathcal{H}_1$, the estimation procedure was the same for all three methods, as no constraints needed to be evaluated.

	M_{uEP}	SD_{uEP}	$\bar{t}_{uEP}$	M_{cEP}	SD_{cEP}	$\bar{t}_{cEP}$	M_{SICS}	SD_{SICS}	$\bar{t}_{SICS}$
$\mathcal{H}_1$	-104.2	0.006	181.7	-104.2	0.006	181.7	-104.2	0.006	181.7
$\mathcal{H}_2$	-94.5	0.011	231.1	-94.5	0.012	234.1	-94.5	0.008	255.6
$\mathcal{H}_3$	-94.6	0.025	239.9	-94.6	0.026	251.0	-94.6	0.011	296.9
$\mathcal{H}_4$	-95.7	0.043	255.0	-95.7	0.043	291.2	-95.7	0.012	345.1
$\mathcal{H}_5$	-96.0	0.256	253.6	-96.1	0.394	325.9	-96.0	0.014	361.0
$\mathcal{H}_6$	-96.9	0.351	260.9	-97.0	0.239	358.0	-96.9	0.016	379.7
$\mathcal{H}_7$	-98.8	0.821	274.2	-98.8	0.292	403.6	-98.7	0.016	400.3
$\mathcal{H}_8$	-	-	296.2	-100.3	0.378	453.3	-100.2	0.016	415.4

TABLE 6.D.2: Efficiency indices, with $\text{eff}_{\text{method}} = SD^{-1}_{\text{method}}/\bar{t}_{\text{method}}$. uEP: unconditional encompassing prior method. cEP: conditional encompassing prior method. SICS: Sampling in constrained space. -: no estimation possible, because no valid marginal likelihood estimates were produced. *: No estimation possible, because SD of uEP could not be estimated. For $\mathcal{H}_1$, the estimation procedure was the same for all three methods, as no constraints needed to be evaluated.

	eff_{uEP}	eff_{cEP}	eff_{SICS}	$\frac{\text{eff}_{\text{cEP}}}{\text{eff}_{\text{uEP}}}$	$\frac{\text{eff}_{\text{SICS}}}{\text{eff}_{\text{uEP}}}$	$\frac{\text{eff}_{\text{SICS}}}{\text{eff}_{\text{cEP}}}$
$\mathcal{H}_1$	0.887	0.887	0.887	1.0	1.0	1.0
$\mathcal{H}_2$	0.382	0.346	0.462	0.9	1.2	1.3
$\mathcal{H}_3$	0.166	0.151	0.297	0.9	1.8	2.0
$\mathcal{H}_4$	0.091	0.080	0.236	0.9	2.6	3.0
$\mathcal{H}_5$	0.015	0.008	0.195	0.5	12.6	25.0
$\mathcal{H}_6$	0.011	0.012	0.161	1.1	14.8	13.8
$\mathcal{H}_7$	0.004	0.008	0.152	1.9	34.2	17.9
$\mathcal{H}_8$	-	0.006	0.147	*	*	25.1

TABLE 6.D.3: Evidence for the eight tested hypotheses, for each of the three models. Presented are log marginal likelihoods. mean con.: Ordinal constraints on the population mean parameters of learning rates. full con.: Ordinal constraints on every single participant learning rate vector. free: No constraints.

Model	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$	$\mathcal{H}_7$	$\mathcal{H}_8$
mean con.	-1620.0	-1524.9	-1480.9	-1462.5	-1424.3	-1416.3	-1426.1	-1432.1
full con.	-1620.0	-1526.3	-1501.1	-1500.3	-1477.8	-1475.0	-1512.2	-1501.9
free	-1620.0	-1525.5	-1481.5	-1464.7	-1427.6	-1422.3	-1427.1	-1434.6

Part III

Conclusion

Chapter 7

General Discussion

7.1 Summary

Science is, ultimately, about the continuous updating of our knowledge in face of new data. Bayesian inference provides a principled framework to do this, and its continued uptake in psychology and social neuroscience promises great advances in the way we reason about our hypotheses, data, and results. However, feasible Bayesian analyses are a relatively "young" addition to neuro-scientific methodology, and therefore still pose many challenges for the empirical researcher. Therefore, the aim of my thesis was to demonstrate how Bayesian methods can be applied to the investigation of social behavior and their neural bases, and to explore the ways Bayesian thinking can improve the scientific process itself.

In **Part I** of this thesis, I applied Bayesian reasoning to the inferences drawn from standard experimental designs. In Chapter 2, I used a fully Bayesian analysis approach to test whether or not violent video games numb their players to the pain of other individuals. Here, the Bayesian approach was particularly well suited, as previous results had been highly inconclusive regarding the presence or absence of such effects. The results suggested that continued exposure to video game violence does not decrease individuals' empathy for pain and emotional reactivity - at least not in tightly controlled experimental settings. However, it remains to be shown if common concerns about the negative effects of violent media are really unfounded, or if negative effects arise after more extensive amounts of exposure, or in particularly vulnerable groups.

In Chapter 3, I discussed the common notion that significant results from underpowered tests provide very little evidence for the tested hypothesis. Shedding light on this question from both a frequentist and Bayesian perspective, I came to the conclusion that low power alone is rarely reason enough to dismiss significant results. On the one hand, assessing the posterior probability of a hypothesis after a significant result is inconsistent with the principles of frequentist inference. On the other hand, even low-powered tests can provide a non-negligible amount of evidence for the tested hypothesis if a Bayesian perspective is adopted. Most importantly, however, questions of power should not distract from questions of scientific integrity, as questionable research practices greatly diminish the evidence provided by low- and high-powered studies alike.

In **Part II**, I focused on the Bayesian estimation of cognitive models, in particular reinforcement learning (RL) models, in social neuroscience. In Chapter 4, my co-authors and I discussed common pitfalls in the application of RL models, and developed suggestions for best practices. Using simulations, we demonstrated that

the interpretation of the learning rate parameter in terms of optimal behavior is rarely straightforward, but depends on the design of the task, as well as other model parameters. Moreover, we illustrated approaches to detect true prediction error signals within the brain, and explained the advantages of Hierarchical Bayesian modeling in research employing RL models.

In Chapter 5, I presented an empirical study on the neural bases of prosocial learning and decision making. I found that participants made more optimal choices when they could avoid harmful consequences (painful shocks) for another person, compared to when they could avoid harmful consequences for themselves. Using Hierarchical Bayesian RL modeling, I further found that this altruistic tendency was due to greater value-sensitivity in prosocial contexts: when participants made choices that could affect the other person, they were more sensitive to the learned values of choice alternatives. On the neural level, social choices were characterized by increased connectivity between brain areas associated with valuation, and brain areas associated with self-other distinction and perspective taking. Overall, these results support the theory that, in contexts of harm-avoidance, humans are "intuitively prosocial" (Zaki & Mitchell, 2013). Moreover, prosocial learning appears to arise from an interaction of brain regions underpinning learning and social cognition.

Finally, in Chapter 6, I developed Sampling In Constrained Space (SICS), an efficient approach for the Bayesian testing of constrained hypotheses, and demonstrated that SICS is able to outperform established algorithms in terms of efficiency and precision. Moreover, as an exemplary analysis, I used SICS to test whether or not human reinforcement learning is characterized by decreasing learning rates. The results provided very strong evidence for decreasing learning rates: at least in the tested sample, participants updated their subjective value about a choice option less the more information they had already collected about it.

With this collection of papers on a wide array of theoretical aspects and applications, I hope to have demonstrated the value of Bayesian inference for neuro-scientific research. In my opinion, however, there are still several challenges that arise with the use of Bayesian inference in social neuroscience, and I would like to use this final part of the thesis to shortly discuss these.

7.2 The Future of Bayesian Inference in a (Still) Frequentist Field

Going forward, should we be frequentists, Bayesians - or both? Numerous publications advertise the advantages of Bayesian inference over traditional frequentist tests, and propose that Bayesian measures of evidence such as the Bayes factor should complement, or even replace, p-values and significance tests (Benjamin et al., 2018; Coventry & Bartlett, 2024; Dienes, 2016; Dienes & Mclatchie, 2018; Keysers et al., 2020; McElreath, 2018; Wagenmakers et al., 2011, 2018). Clearly, there is great potential in the increased use of Bayesian methods in neuroscience and other research fields. At the same time, the advantages of Bayesian inference come at the 'cost' of implicitly or explicitly accepting a set of assumptions and goals that are, in part, inconsistent with the way inference has been performed so far. Without a critical and open discussion about which aims statistical inference *should* fulfill, and which approach best fulfills that aim, research runs the risk of becoming arbitrary in its definition of what constitutes 'good' statistical practice.

Bayesian and 'classical' frequentist inference are based on different conceptions of probability, and designed to achieve different goals (see also Chapter 3). Bayesian inference accepts a subjective definition of probability (i.e., probability as a quantification of belief or knowledge), and aims at the optimal updating of belief (represented by prior probabilities and distributions) in the view of data (De Finetti, 1974; Jaynes, 2003). In line with this, statements about the probability of a hypothesis being 'true' are considered meaningful, inasmuch as equating probability with belief is considered valid. In contrast, frequentist inference explicitly rejects subjective probability, and proposes that probability may only be understood as the long-run frequency of repeatable events (Fisher, 1935; Neyman & Pearson, 1933). Consequently, frequentist inference is explicitly agnostic about the truth or falsehood of hypotheses. Rather, its aim is to achieve optimal error control in the long run of decisions[1]: A 'good' frequentist analysis is one that tightly controls the Type-I error rate (when the null hypothesis is true, it is rejected only in a prespecified percentage of cases, typically 5%), and minimizes the Type-II error rate respectively maximizes the power (when the alternative hypothesis is true, it is accepted as often as possible).

7.2.1 The Case of Testing for 'Evidence of Absence'

Philosophical nuances in the interpretation of probability might be of little concern to the empirical researcher. Fundamental differences in the goals of inference should not be, though, as they have direct consequences for the criteria of 'good' statistics.

Consider, for example, the following scenario involving the use of Bayesian inference to provide evidence for the 'null': A researcher wants to test two related research hypotheses. Conscientious and knowledgeable about current best practices, they set the desired type I error level to the conventional 5%, conduct a power analysis, preregister the planned analyses, collect the required samples, compute only the preregistered significance tests, and correct for multiple comparisons where necessary. The researcher finds a significant result for one hypothesis (say, $p = 0.028$), but no significance for the other. Evidence for the absence of the second hypothesized effect would be an interesting addition to the paper, but the researcher also knows that an insignificant result does not represent 'evidence of absence'. Therefore, they follow popular suggestions (e.g. Dienes, 2014; Keysers et al., 2020) and compute the

[1]This decision-oriented framework is certainly useful in applications of statistics (e.g., quality control, diagnostic screening) where a test can be followed up with a practically meaningful decision (e.g., throw out a product, or consider a patient for further medical screening). However, in scientific inference, it is often not clear what practically relevant 'decision' a scientist should make after a significant or insignificant result. Arguably, it should be the decision to 'accept' or 'reject' the hypothesis. But, as Rozeboom (1960) writes poignantly,

> ... a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true (p. 423).

And a clear-cut acceptance or rejection of the hypothesis is also rarely done in practice:

> If we did, in fact, accept or reject the null hypothesis according to whether the sample statistic falls in the acceptance or in the rejection region, then there would be no replications of experimental designs, no multiplicity of experimental approaches to an important hypothesis—a single experiment would, by definition of the method, make up our mind about the hypothesis in question. And the fact that in actual practice, a single finding seldom even tempts us to such closure of judgment reveals how little the conventional model of hypothesis testing fits our actual evaluative behavior (Rozeboom, 1960, p. 420).

Bayes factor to test if the data really provides evidence for the null hypothesis. The researcher finds a Bayes factor smaller than 1/3, and writes up a paper about two findings: evidence for the presence of the first effect, and evidence for the absence of the second.

Up until the calculation of the Bayes factor, the researcher acts perfectly, and recommendably, in line with best practices informed by frequentist principles: They attempt to minimize type II errors by conducting a power analysis, they make sure to properly control the type I error by correcting for multiple comparisons, and they minimize the detrimental effects of researcher degrees-of-freedom (Wicherts et al., 2016) by following a preregistered analysis plan. By implementing all of these measures, the researcher makes sure that their inference adheres to the frequentist ideal of tight error control, and typical reviewers would rightfully be impressed by their methodological thoroughness. However, when the researcher calculates the Bayes factor to support the null hypothesis, they suddenly switch to an analysis framework that does not satisfy the criteria they meticulously followed throughout the rest of their study. Bayes factors are not designed for error control. The researcher would need to conduct a non-trivial additional simulation experiment (Schönbrodt & Wagenmakers, 2018) to assess how often, given the design and sample size, they would accept the null hypothesis if it were true - and unless the study were explicitly designed to achieve such 'Bayesian error control', the numbers would likely be far of from the type I and II error rates originally considered. Indeed, when the researcher follows typical guidelines and interprets Bayes factors between 1/3 and 3 as inconclusive evidence for either hypothesis, they introduce a third decision ('neither accept or reject either hypothesis') that does not even fit into the strictly binary decision framework of frequentist inference.

In effect, our exemplary researcher 'accepted' the two hypotheses based on fundamentally different criteria. They 'accepted' the presence of the first effect because this was dictated by the decision procedure that, in the long run, minimizes the rate of false decisions. They 'accepted' the absence of the second effect because the data were better predicted under the assumption of no effect than otherwise. Thus, switching from a p-value to a Bayes factor does not simply mean using a 'more advanced' or 'better' statistic - it means asking a different question.

7.2.2 The Case of Optional Stopping

Admittedly, using Bayes factors to support the null hypothesis is a relatively mild case of inconsistency between the frequentist and Bayesian framework. A study that otherwise fulfills frequentist quality criteria can report Bayes factors as additional information, to be considered or ignored at the discretion of the reader. Greater issues arise when one framework allows or even encourages procedures that are explicitly rejected as bad practice in the other. Perhaps the most illustrative example is *optional stopping*, the collection of samples until some decision threshold is reached (De Heide & Grünwald, 2021; Rouder, 2014).

Consider, again, two hypothetical scenarios, involving a researcher who wants to test some hypothesis. In scenario 1, the researcher collects an initial sample, and performs a significance test. After the test is not significant ($p > 0.05$), the researcher collects more samples, and, after each new incoming data point, recalculates the p-value. Ultimately, the test produces a p-value < 0.05, and the researcher writes a paper in which they presents this as evidence for their hypothesis.

Now, if the paper even makes it past the editorial desk, it is bound to be taken apart by reviewers[2]. Clearly, the large number of tests the researcher conducted on the same data vastly inflated the type I error rate. For the result to have even a slight chance of being considered evidence, it would need to survive extensive correction for multiple comparisons.

In scenario 2, the researcher does exactly the same - but instead of p-values, they compute Bayes factors until some threshold of evidence for the hypothesis, such as $BF > 3$, is reached. When presenting the results, the researcher may now refer to numerous publications that assert that optional stopping is allowed in Bayesian inference, and even encouraged to ensure an efficient use of resources (e.g. Dienes, 2016; Keysers et al., 2020; Rouder, 2014; Wagenmakers et al., 2012). Consequently, reviewers and other readers are far more likely to assess the result as credible evidence.

What changed between scenario 1 and scenario 2? Did the intention of the researcher or the quality of the data change? Arguably not. Did the researcher of scenario 2 use a method that is immune to the effects of alpha error cumulation? No - the Bayes factor is a statistic subject to sampling variation just like the p-value, and may thus reach the decision threshold by random chance alone. Calculating the BF at multiple time points inevitably increases the probability of observing one that is larger than the threshold (De Heide & Grünwald, 2021; Sanborn & Hills, 2014; Yu et al., 2014). If error control is seen as the goal of statistical inference, then Bayesian hypothesis testing under optional stopping does not reach it.

However, when the researcher reports Bayes factors instead of p-values, they signal to reviewers and readers that error control is *not* the aim of their inference. It is then up to the evaluators of their work to decide whether this departure from frequentist principles is acceptable. In practice, unless particularly committed to frequentist doctrine, evaluators are often swayed by the growing number of publications asserting that optional stopping is "okay if it's Bayesian." Thus, practices that would typically be criticized as serious methodological flaws are suddenly tolerated - not because the underlying standards have changed, but because they are temporarily suspended under the Bayesian label.

7.2.3 Towards Inference we Agree on

Given the increasing number of publications employing Bayesian inference today, it appears that the scientific community is, in practice, willing to accept modes of inference that forego the frequentist principle of error control. At the same time, this very principle still underpins many influential decisions in the scientific process: Papers are accepted or rejected, grants are awarded or denied, based on researchers' adherence to frequentist quality criteria. A large amount of theoretical papers warn scientists about the detrimental effects of undisclosed multiple testing and other researcher degrees of freedom, and in Statistics courses, students are still taught that tight error control is a key aspect of 'good' inference. It is somewhat surprising, thus, how quickly these principles are set aside as soon as an analysis is called 'Bayesian'.

From everything I have written in this thesis, it should be obvious that I am far from opposed to the continued use of Bayesian inference in social neuroscience and psychology. On the contrary, I believe it holds great promise. However, in my

[2]That is, if the researcher was conscientious enough to transparently report their sampling procedure in the paper. But, as research on questionable research practices has shown, it would not be entirely untypical for the researcher to omit this information in the submitted manuscript (e.g. John et al., 2012).

opinion, we as a scientific community owe it to ourselves, and to the integrity of the scientific process, to reflect and openly discuss what goals we wish to achieve with statistical inference, and which inferential framework best aligns with those goals. It is unlikely that the result of this discussion will be a clear-cut decision as to whether our field is 'frequentist' or 'Bayesian'. Both approaches are useful in different situations, and often, their results converge to the same conclusion. When they disagree, however, only careful consideration of our goals can tell us which inferential framework provides the evidence we wish to base future research on.

Annex

Bibliography

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2015). Computational Psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery and Psychiatry*, jnnp–2015-310737. https://doi.org/10.1136/jnnp-2015-310737

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, *1*(Figure 1), 24–57. https://doi.org/10.1162/CPSY_a_00002

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected papers of Hirotogu Akaike* (pp. 199–213).

Altman, D. G., & Bland, j. M. (1994). Statistics Notes: Diagnostic tests 2: Predictive values. *BMJ*, *309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R., & Saleem, M. (2010). Violent Video Game Effects on Aggression, Empathy, and Prosocial Behavior in Eastern and Western Countries: A Meta-Analytic Review. *Psychological Bulletin*, *136*(2), 151–173. https://doi.org/10.1037/a0018251

Arriaga, P., Monteiro, M. B., & Esteves, F. (2011). Effects of Playing Violent Computer Games on Emotional Desensitization and Aggressive Behavior. *Journal of Applied Social Psychology*, *41*(8), 1900–1925. https://doi.org/10.1111/j.1559-1816.2011.00791.x

Ashar, Y. K., Andrews-Hanna, J. R., Dimidjian, S., & Wager, T. D. (2017). Empathic care and distress: Predictive brain markers and dissociable brain systems. *Neuron*, *94*(6), 1263–1273.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Aylward, J., Valton, V., Ahn, W. Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature Human Behaviour*, *3*. https://doi.org/10.1038/s41562-019-0628-0

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on psychological science*, *7*(6), 543–54. https://doi.org/10.1177/1745691612459060

Bartholow, B. D., & Anderson, C. A. (2002). Effects of Violent Video Games on Aggressive Behavior: Potential Sex Differences. *Journal of Experimental Social Psychology*, *38*(3), 283–290. https://doi.org/10.1006/jesp.2001.1502

Bartholow, B. D., Bushman, B. J., & Sestir, M. A. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. *Journal of Experimental Social Psychology*, *42*(4), 532–539. https://doi.org/10.1016/j.jesp.2005.08.006

Bartholow, B. D., Sestir, M. A., & Davis, E. B. (2005). Correlates and consequences of exposure to video game violence: Hostile personality, empathy, and aggressive behavior. *Personality and Social Psychology Bulletin*, *31*(11), 1573–1586. https://doi.org/10.1177/0146167205277205

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*(1), 132–150. https://doi.org/10.1016/j.jmp.2013.12.002

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batteux, E., Ferguson, E., & Tunney, R. J. (2019). Do our risk preferences change when we make decisions for others? a meta-analysis of self-other differences in decisions involving risk. *PloS one*, *14*(5), e0216566.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*(2), 142–163. https://doi.org/10.1037/1082-989X.11.2.142

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

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

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature human behaviour*, *2*(1), 6–10.

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*(1), 1–32. https://doi.org/10.1214/ss/1056397485

Berger, J. O., & Mortera, J. (1999). Default bayes factors for nonnested hypothesis testing. *Journal of the american statistical association*, *94*(446), 542–554.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. https://doi.org/10.48550/ARXIV.1701.02434

Bettencourt, B. A., & Kernahan, C. (1997). A meta-analysis of aggression in the presence of violent cues: Effects of gender differences and aversive provocation. *Aggressive Behavior*, *23*(6), 447–456. https://doi.org/10.1002/(SICI)1098-2337(1997)23:6<447::AID-AB4>3.0.CO;2-D

Bird, G., Silani, G., Brindley, R., White, S., Frith, U., & Singer, T. (2010). Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain : a journal of neurology*, *133*(Pt 5), 1515–1525. https://doi.org/10.1093/brain/awq060

Borghi, O., Mayrhofer, L., Voracek, M., & Tran, U. S. (2023). Differential associations of the two higher-order factors of mindfulness with trait empathy and the mediating role of emotional awareness. *Scientific Reports*, *13*(1), 3201. https://doi.org/10.1038/s41598-023-30323-6

Borkenau, P., & Ostendorf, F. (1993). *Neo-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae : Handanweisung*. Hogrefe.

Bottemanne, H. (2024). Bayesian brain theory: Computational neuroscience of belief. *Neuroscience*, *566*, 198–204. https://doi.org/10.1016/j.neuroscience.2024.12.003

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/JSS.V080.I01

Bushman, B. J. (1998). Priming Effects of Media Violence on the Accessibility of Aggressive Constructs in Memory. *Personality and Social Psychology Bulletin*, *24*(5), 537–545. https://doi.org/10.1177/0146167298245009

Bushman, B. J., & Anderson, C. A. (2002). Violent Video Games and Hostile Expectations: A Test of the General Aggression Model. *Personality and Social Psychology Bulletin*, *28*(12), 1679–1686. https://doi.org/10.1177/014616702237649

Bushman, B. J., & Anderson, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science*, *20*(3), 273–277. https://doi.org/10.1111/j.1467-9280.2009.02287.x

Bushman, B. J., & Anderson, C. A. (2015). Understanding Causality in the Effects of Media Violence. *American Behavioral Scientist*, *59*(14), 1807–1821. https://doi.org/10.1177/0002764215596554

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Calvert, S. L., Appelbaum, M., Dodge, K. A., Graham, S., Nagayama Hall, G. C., Hamby, S., Fasig-Caldwell, L. G., Citkowicz, M., Galloway, D. P., & Hedges, L. (2017). The american psychological association task force assessment of violent video games: Science in the service of public interest. *American Psychologist*, *72*(2), 126–143. https://doi.org/10.1037/a0040413

Carnagey, N. L., Anderson, C. A., & Bushman, B. J. (2007). The effect of video game violence on physiological desensitization to real-life violence. *Journal of Experimental Social Psychology*, *43*. https://doi.org/10.1016/j.jesp.2006.05.003

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Carter, R. M. K., & Huettel, S. A. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, *17*(7), 328–336. https://doi.org/10.1016/j.tics.2013.05.007

Chen, C., Martínez, R. M., & Cheng, Y. (2018). The developmental origins of the social brain: Empathy, morality, and justice. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02584

Chen, Y., Schomaker, L., & Wiering, M. (2021). An Investigation Into the Effect of the Learning Rate on Overestimation Bias of Connectionist Q-learning. *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. https://doi.org/10.5220/0010227301070118

Chib, S., & Jeliazkov, I. (2001). Marginal Likelihood From the Metropolis–Hastings Output. *Journal of the American Statistical Association*, *96*(453), 270–281. https://doi.org/10.1198/016214501750332848

Chien, S., Wiehler, A., Spezio, M., & Glascher, J. (2016). Congruence of Inherent and Acquired Values Facilitates Reward-Based Decision-Making. *Journal of Neuroscience*, *36*(18), 5003–5012. https://doi.org/10.1523/JNEUROSCI.3084-15.2016

Chung, D., Kadlec, K., Aimone, J. A., McCurry, K., King-Casas, B., & Chiu, P. H. (2017). Valuation in major depression is intact and stable in a non-learning environment. *Scientific Reports*, *7*. https://doi.org/10.1038/srep44374

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43245-z

Clement, J. (2021). Weekly hours spent playing video games worldwide 2021, by country [[Online; accessed 2023-08-29]]. https://www.statista.com/statistics/273829/average-game-hours-per-day-of-video-gamers-in-selected-countries/

Cohen, J. (2013). *Statistical power analysis for the Behavioral Sciences*. Academic Press.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, *20*(3), 304–313. https://doi.org/10.1038/nn.4499

Colling, L. J., & Szűcs, D. (2021). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*, *12*(1), 121–147.

Collins, A. G. E., & Shenhav, A. (2021). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, *47*(1), 104–118. https://doi.org/10.1038/s41386-021-01126-y

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p -values. *Royal Society Open Science*, *1*(3), 140216. https://doi.org/10.1098/rsos.140216

Cools, R., Altamirano, L., & D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. *Neuropsychologia*, *44*(10), 1663–1673. https://doi.org/10.1016/j.neuropsychologia.2006.03.030

Coventry, B. S., & Bartlett, E. L. (2024). Practical bayesian inference in neuroscience: Or how i learned to stop worrying and embrace the distribution. *Eneuro*, *11*(7).

Crawley, D., Zhang, L., Jones, E. J., Ahmad, J., Oakley, B., San José Cáceres, A., Charman, T., Buitelaar, J. K., Murphy, D. G., Chatham, C., et al. (2020). Modeling flexible behavior in childhood to adulthood shows age-dependent learning mechanisms and less optimal learning in autism in each age group. *PLoS biology*, *18*(10), e3000908.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. https://doi.org/10.1016/j.tics.2013.06.005

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(40), 17433–17438. https://doi.org/10.1073/pnas.1009396107

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, *320*(5884), 1739. https://doi.org/10.1126/science.1155577

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(48), 17320–5. https://doi.org/10.1073/pnas.1408988111

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable Effects of Serotonin and Dopamine on the Valuation of Harm in Moral Decision Making. *Current Biology*, *25*(14), 1852–1859. https://doi.org/10.1016/j.cub.2015.05.021

Daunizeau, J., Adam, V., & Rigoux, L. (2014). Vba: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*, *10*(1). https://doi.org/10.1371/journal.pcbi.1003441

Davison, A. C., & Hinkley, D. (1997, October). *Bootstrap Methods and their Application*. Cambridge University Press. https://doi.org/10.1017/cbo9780511802843

Daw, N. D., O'Doherty, J. P., Dayan, P., Dolan, R. J., & Seymour, B. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–9. https://doi.org/10.1038/nature04766

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and Performance XXIII*, 1–26. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

Daw, N. D. (2013). Advanced Reinforcement Learning. *Neuroeconomics: Decision Making and the Brain: Second Edition*, 299–320. https://doi.org/10.1016/B978-0-12-416008-8.00016-4

de Vrieze, J. (2018). The metawars. *Science*, *361*(6408).

De Finetti, B. (1974). *Theory of probability*. Wiley.

De Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin & Review*, *28*, 795–812.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences*, *22*(9), 764–779.

DeCamp, W., & Ferguson, C. J. (2016). The Impact of Degree of Exposure to Violent Video Games, Family Background, and Other Factors on Youth Violence. *Journal of Youth and Adolescence 2016 46:2*, *46*(2), 388–400. https://doi.org/10.1007/S10964-016-0561-8

Decety, J., Bartal, I. B. A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: Highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686). https://doi.org/10.1098/rstb.2015.0077

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology*, *30*(1), 153–164. https://doi.org/10.1017/S0954579417000530

Den Ouden, H. E., Daw, N. D., Fernandez, G., Elshout, J. A., Rijpkema, M., Hoogman, M., Franke, B., & Cools, R. (2013). Dissociable Effects of Dopamine and Serotonin on Reversal Learning. *Neuron*, *80*(4), 1090–1100. https://doi.org/10.1016/j.neuron.2013.08.030

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.

Dienes, Z. (2016). How bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dienes, Z., Coulton, S., & Heather, N. (2018). Using bayes factors to evaluate evidence for no effect: Examples from the sips project. *Addiction*, *113*(2), 240–246.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207–218.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doshi, P., Zeng, Y., & Chen, Q. (2009). Graphical models for interactive POMDPs: Representations and solutions. *Autonomous Agents and Multi-Agent Systems*, *18*(3), 376–416. https://doi.org/10.1007/s10458-008-9064-7

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*(397), 171. https://doi.org/10.2307/2289144

Eisenegger, C., Pedroni, A., Rieskamp, J., Zehnder, C., Ebstein, R., Fehr, E., & Knoch, D. (2013). Dat1 polymorphism determines l-dopa effects on learning about others' prosociality. *PLoS One*, *8*(7), e67820.

Eldar, E., Hauser, T. U., Dayan, P., & Dolan, R. J. (2016). Striatal structure and function predict individual biases in learning to avoid pain. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(17), 4812–4817. https://doi.org/10.1073/pnas.1519829113

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis [doi: 10.1177/0956797620916786]. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Engelhardt, C. R., Bartholow, B. D., Kerr, G. T., & Bushman, B. J. (2011). This is your brain on violent video games: Neural desensitization to violence predicts increased aggression following violent video game exposure. *Journal of Experimental Social Psychology*, *47*(5), 1033–1036. https://doi.org/10.1016/j.jesp.2011.03.027

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, *1*(4), 429–438. https://doi.org/10.3758/BF03210947

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects in eastern and western nations: Comment on Anderson et al. (2010). *Psychological bulletin*, *136*(2), 174–187. https://doi.org/10.1037/a0018566

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., Hunt, L. T., & Corlett, P. R. (2018). Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biological Psychiatry*, 1–8. https://doi.org/10.1016/j.biopsych.2018.05.020

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T., Cavanagh, J. F., & Badre, D. (2015). Fmri and EEG Predictors of Dynamic Decision Parameters during

Human Reinforcement Learning. *Journal of Neuroscience*, *35*(2), 485–494. https://doi.org/10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221. https://doi.org/10.1098/rstb.2008.0300

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense [doi: 10.1177/2515245919847202]. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(33), E7680–E7689. https://doi.org/10.1073/pnas.1802523115

Gao, X., Pan, W., Li, C., Weng, L., Yao, M., & Chen, A. (2017). Long-time exposure to violent video games does not show desensitization on empathy for pain: An fMRI Study. *Frontiers in Psychology*, *8*, 650. https://doi.org/10.3389/fpsyg.2017.00650

Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 1682–1690. http://proceedings.mlr.press/v84/ge18b.html

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013, January). *Bayesian data analysis, third edition*. CRC Press. https://doi.org/10.1201/b16018

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in bayesian social research. *Sociological methodology*, *25*, 165–173.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence*, *27*(1), 5–22. https://doi.org/10.1016/j.adolescence.2003.10.002

Gentile, D. A., Swing, E. L., Anderson, C. A., Rinker, D., & Thomas, K. M. (2016). Differential neural recruitment during violent video game play in violent- and nonviolent-game players. *Psychology of Popular Media Culture*, *5*(1), 39–51. https://doi.org/10.1037/ppm0000009

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin and Review*, *22*(5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3

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

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, *19*, 483–495. https://doi.org/10.1093/cercor/bhn098

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: Combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 501–510. https://doi.org/10.1002/wcs.57

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The bayesian two-sample t test. *The American Statistician*, *59*(3), 252–257.

Gorinova, M., Moore, D., & Hoffman, M. (2020). Automatic reparameterisation of probabilistic programs. *International Conference on Machine Learning*, 3648–3657.

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and social psychology bulletin*, *40*(5), 578–589.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Grueschow, M., Polania, R., Hare, T. A., & Ruff, C. C. (2015). Automatic versus Choice-Dependent Value Representations in the Human Brain. *Neuron*, *85*(4), 874–885. https://doi.org/10.1016/j.neuron.2014.12.054

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*(8), 1526–1553.

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J., Dayan, P., Dolan, R. J., & Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *Journal of Neuroscience*, *31*(21), 7867–7875. https://doi.org/10.1523/JNEUROSCI.6376-10.2011

Guo, X., Zheng, L., Wang, H., Zhu, L., Li, J., Wang, Q., Dienes, Z., & Yang, Z. (2013). Exposure to violence reduces empathetic responses to other's pain. *Brain and Cognition*, *82*(2), 187–191. https://doi.org/10.1016/j.bandc.2013.04.005

Haaf, J. M., Merkle, E. C., & Rouder, J. N. (2020). Do items order? The psychology in IRT models. *Journal of Mathematical Psychology*, *98*, 102398. https://doi.org/10.1016/j.jmp.2020.102398

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Haaf, J. M., & Rouder, J. N. (2023). Does every study? Implementing ordinal constraint in meta-analysis. *Psychological Methods*, *28*(2), 472–487. https://doi.org/10.1037/met0000428

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001

Haines, N., Vassileva, J., & Ahn, W. Y. (2018). The Outcome-Representation Learning Model: A Novel Reinforcement Learning Model of the Iowa Gambling Task. *Cognitive Science*, *42*(8), 2534–2561. https://doi.org/10.1111/cogs.12688

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, *156*, 563–585.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(18), 6741–6746. https://doi.org/10.1073/pnas.0711099105

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, *30*(2), 583–590. https://doi.org/10.1523/JNEUROSCI.4089-09.2010

Hartmann, H., Forbes, P. A. G., Rütgen, M., & Lamm, C. (2022). Placebo analgesia reduces costly prosocial helping to lower another person's pain. *Psychological Science*, *33*(11), 1867–1881.

Hartmann, H., Rütgen, M., Riva, F., & Lamm, C. (2021). Another's pain in my brain: No evidence that placebo analgesia affects the sensory-discriminative component in empathy for pain. *NeuroImage*, *224*, 117397. https://doi.org/10.1016/J.NEUROIMAGE.2020.117397

Hasan, Y., Bègue, L., Scharkow, M., & Bushman, B. J. (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. *Journal of Experimental Social Psychology*, *49*(2), 224–227. https://doi.org/10.1016/j.jesp.2012.10.016

Hauser, T. U., Iannaccone, R., Walitza, S., Brandeis, D., & Brem, S. (2015). Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. *NeuroImage*, *104*, 347–354.

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for bayesian inference. *Journal of mathematical psychology*, *91*, 70–87.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron*, *68*(1), 149–160. https://doi.org/10.1016/j.neuron.2010.09.003

Hein, G., & Singer, T. (2008). I feel how you feel but not always: The empathic brain and its modulation. *Current opinion in neurobiology*, *18*(2), 153–158. https://doi.org/10.1016/j.conb.2008.07.012

Hilgard, J., Engelhardt, C. R., Bartholow, B. D., & Rouder, J. N. (2017). How much evidence is p >.05? Stimulus pre-testing and null primary outcomes in violent video games research. *Psychology of Popular Media Culture*, *6*(4), 361–380. https://doi.org/10.1037/ppm0000102

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson

et al. (2010). *Psychological Bulletin*, *143*(7), 757–774. https://doi.org/10.1037/bul0000074

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, *20*(8), 1142–1149. https://doi.org/10.1038/nn.4602

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Hoijtink, H. (2022). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods*, *27*(5), 804.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J. C., & Weber, B. (2018). Spreading inequality: Neural computations underlying paying-it-forward reciprocity. *Social Cognitive and Affective Neuroscience*, *13*(6), 578–589. https://doi.org/10.1093/scan/nsy040

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLoS Computational Biology*, *14*(2), 1–20. https://doi.org/10.1371/journal.pcbi.1005935

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, 9. https://doi.org/10.3389/fnins.2012.00009

Ihssen, N., Mussweiler, T., & Linden, D. E. (2016). Observing others stay or switch – How social prediction errors are integrated into reward reversal learning. *Cognition*, *153*, 19–32. https://doi.org/10.1016/j.cognition.2016.04.012

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

IPSOS MediaCT. (2012). Videogames in Europe: Consumer study. http://www.isfe.eu/sites/isfe.eu/files/attachments/euro_summary_-_isfe_consumer_study.pdf

Janowski, V., Camerer, C., & Rangel, A. (2013). Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. *Social Cognitive and Affective Neuroscience*, *8*(2), 201–208. https://doi.org/10.1093/scan/nsr086

JASP Team. (2025). JASP (Version 0.19.3)[Computer software]. https://jasp-stats.org/

Jauniaux, J., Khatibi, A., Rainville, P., & Jackson, P. L. (2019). A meta-analysis of neuroimaging studies on pain empathy: Investigating the role of visual information and observers' perspective. *Social cognitive and affective neuroscience*, *14*(8), 789–813. https://doi.org/10.1093/scan/nsz055

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jocham, G., Furlong, P. M., Kröger, I. L., Kahn, M. C., Hunt, L. T., & Behrens, T. E. (2014). Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. *NeuroImage*, *100*, 498–506.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Powers, A., Mehta, N., Dyke, J., & Casey, B. J. (2014). Adolescent-specific patterns of behavior and neural activity during social reinforcement learning. *Cognitive, affective & behavioral neuroscience*, *14*(2), 683–97. https://doi.org/10.3758/s13415-014-0257-z

Kable, J. W., & Glimcher, P. W. (2009). The Neurobiology of Decision: Consensus and Controversy. *Neuron*, *63*(6), 733–745. https://doi.org/10.1016/j.neuron.2009.09.003

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kamary, K., Mengersen, K., Robert, C. P., & Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. *arXiv*, 1412.2044.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, *66*, 59–69. https://doi.org/10.1016/j.jmp.2015.03.006

Keller, M., & Kamary, K. (2017). Bayesian Model Averaging By Mixture Modeling. *arXiv*, 1711.10016v2.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological methods*, *8*(2), 115–28. https://doi.org/10.1037/1082-989x.8.2.115

Keum, S., & Shin, H.-S. (2019). Neural Basis of Observational Fear Learning: A Potential Model of Affective Empathy. *Neuron*, *104*(1), 78–86. https://doi.org/10.1016/j.neuron.2019.09.013

Keysers, C., Gazzola, V., & Wagenmakers, E. J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. https://doi.org/10.1038/s41593-020-0660-4

Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature communications*, *8*(1), 16033.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–51. https://doi.org/10.1016/j.neuron.2008.11.027

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological methods*, *10*(4), 477.

Koizumi, A., Amano, K., Cortese, A., Shibata, K., Yoshida, W., Seymour, B., Kawato, M., & Lau, H. (2017). Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*, *1*(1), 1–7. https://doi.org/10.1038/s41562-016-0006

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*, *24*, 41–47. https://doi.org/10.1016/j.copsyc.2018.04.009

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(1), 325. https://doi.org/10.1038/s41467-017-02750-3

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020) [doi: 10.1177/0956797621989730]. *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

Krahe, B., Moller, I., Kirwil, L., Huesmann, L. R., Felber, J., & Berger, A. (2011). Desensitization to Media Violence: Links With Habitual Media Violence Exposure,

Aggressive Cognitions, and Aggressive Behavior. *Journal of Personality and Social Psychology*, *100*, 630–646. https://doi.org/10.1037/a0021711

Krahé, B., & Möller, I. (2004). Playing violent electronic games, hostile attributional style, and aggression-related norms in German adolescents. *Journal of Adolescence*, *27*(1), 53–69. https://doi.org/10.1016/j.adolescence.2003.10.006

Krantz, A., Shukla, V., & Knox, M. (2017). Violent video games exposed: A blow by blow account of senseless violence in games. *The Journal of Psychology*, *151*(1), 76–87.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kühn, S., Kugler, D., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). The Myth of Blunted Gamers: No Evidence for Desensitization in Empathy for Pain after a Violent Video Game Intervention in a Longitudinal fMRI Study on Non-Gamers. *NeuroSignals*, *26*(1), 22–30. https://doi.org/10.1159/000487217

Kulkarni, S., & Gilbert, H. (2011). The optimal bayes decision rule. In *An Elementary Introduction to Statistical Learning Theory* (pp. 45–54). John Wiley & Sons, Inc.

Kutlikova, H. H., Zhang, L., Eisenegger, C., van Honk, J., & Lamm, C. (2023). Testosterone eliminates strategic prosocial behavior through impacting choice consistency in healthy males. *Neuropsychopharmacology*, *48*(10), 1541–1550.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). Lmertest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/JSS.V082.I13

Kwak, Y., Pearson, J., & Huettel, S. A. (2014). Differential Reward Learning for Self and Others Predicts Self-Reported Altruism. *PLoS ONE*, *9*(9). https://doi.org/10.1371/JOURNAL.PONE.0107621

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lamm, C., Bukowski, H., & Silani, G. (2016). From shared to distinct self-other representations in empathy: Evidence from neurotypical function and socio-cognitive disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686). https://doi.org/10.1098/rstb.2015.0083

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Lamm, C., Rütgen, M., & Wagner, I. C. (2019). Imaging empathy and prosocial emotions. *Neuroscience Letters*, *693*, 49–53. https://doi.org/10.1016/j.neulet.2017.06.054

Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (2005). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL.

Lebreton, M., Bavard, S., Daunizeau, J., & Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0681-8

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, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9. https://doi.org/10.1038/s41562-017-0067

Lemmens, J. S., & Bushman, B. J. (2006). The appeal of violent video games to lower educated aggressive adolescent boys from two countries. *Cyberpsychology and Behavior*, *9*(5), 638–641. https://doi.org/10.1089/cpb.2006.9.638

Lengersdorff, L. L., Wagner, I. C., Lockwood, P. L., & Lamm, C. (2020). When Implicit Prosociality Trumps Selfishness: The Neural Valuation System Underpins More Optimal Choices When Learning to Avoid Harm to Others Than to Oneself. *The Journal of Neuroscience*, *40*(38), 7286–7299. https://doi.org/10.1523/jneurosci.0842-20.2020

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior Predictive Model Checking for Multidimensionality in Item Response Theory. *Applied Psychological Measurement*, *33*(7), 519–537. https://doi.org/10.1177/0146621608329504

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/https://doi.org/10.1016/j.jmva.2009.04.008

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature neuroscience*, *14*(10), 1250–1252.

Lin, A., Rangel, A., & Adolphs, R. (2012). Impaired learning of social compared to monetary rewards in autism. *Frontiers in Neuroscience*, *6*, 1–7. https://doi.org/10.3389/fnins.2012.00143

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Lindström, B., Bellander, M., Chang, A., Tobler, P. N., & Amodio, D. M. (2019). A computational reinforcement learning account of social media engagement. *PsyArXiv*, 1–26.

Lindström, B., Golkar, A., Jangard, S., Tobler, P. N., & Olsson, A. (2019). Social threat learning transfers to decision making in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4732–4737. https://doi.org/10.1073/pnas.1810180116

Lindström, B., Haaker, J., & Olsson, A. (2018). A common neural network differentially mediates direct and social fear learning. *NeuroImage*, *167*, 121–129. https://doi.org/10.1016/J.NEUROIMAGE.2017.11.039

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375.

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(35), 9763–9768. https://doi.org/10.1073/pnas.1603198113

Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., & Apps, M. A. (2017). Prosocial apathy for helping others when effort is required. *Nature human behaviour*, *1*(7), 0131.

Lockwood, P. L., & Klein-Flügge, M. (2019). Computational modelling of social cognition and behaviour – a reinforcement learning primer Patricia L. Lockwood. *PsyArXiv*, 1–28. https://doi.org/10.31234/osf.io/r69e7

Lockwood, P. L., Klein-flügge, M., Abdurahman, A., & Crockett, M. J. (2019). Neural signatures of model-free learning when avoiding harm to self and other. *BioRXiv*. https://doi.org/10.1101/718106

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

Lockwood, P. L., & Wittmann, M. K. (2018). Ventral anterior cingulate cortex and social decision-making. *Neuroscience and Biobehavioral Reviews*, *92*, 187–191. https://doi.org/10.1016/j.neubiorev.2018.05.030

Lockwood, P. L., Wittmann, M. K., Apps, M. A., Klein-Flügge, M. C., Crockett, M. J., Humphreys, G. W., & Rushworth, M. F. (2018). Neural mechanisms for learning self and other ownership. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-018-07231-9

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55. https://doi.org/10.1016/j.jmp.2016.01.003

Lynch, S. M., & Western, B. (2004). Bayesian Posterior Predictive Checks for Complex Models. *Sociological Methods and Research*, *32*(3), 301–335. https://doi.org/10.1177/0049124103257303

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

Mathur, M. B., & VanderWeele, T. J. (2019). Finding Common Ground in Meta-Analysis "Wars" on Violent Video Games. *Perspectives on Psychological Science*, *14*(4). https://doi.org/10.1177/1745691619850104

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 1–24. https://doi.org/10.3389/fnhum.2014.00825

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/J.JML.2017.01.001

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Mayo, D. G., & Morey, R. D. (2017). A Poor Prognosis for the Diagnostic Screening Critique of Statistical Tests. https://doi.org/10.31219/osf.io/ps38b

McCulloch, C. E., & Neuhaus, J. M. (2005). Generalized linear mixed models. *Encyclopedia of biostatistics*, *4*.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 831–860.

Mikus, N., Eisenegger, C., Mathys, C., Clark, L., Müller, U., Robbins, T. W., Lamm, C., & Naef, M. (2023). Blocking d2/d3 dopamine receptors in male participants increases volatility of beliefs when learning to trust others. *Nature Communications*, *14*(1), 4049.

Mikus, N., Korb, S., Massaccesi, C., Gausterer, C., Graf, I., Willeit, M., Eisenegger, C., Lamm, C., Silani, G., & Mathys, C. (2022). Effects of dopamine d2/3 and opioid receptor antagonism on the trade-off between model-based and model-free behaviour in healthy volunteers. *Elife*, *11*, e79661.

Mobbs, D., Marchant, J. L., Hassabis, D., Seymour, B., Tan, G., Gray, M., Petrovic, P., Dolan, R. J., & Frith, C. D. (2009). From Threat to Fear: The Neural Organization of Defensive Fear Systems in Humans. *The Journal of Neuroscience*, *29*(39), 12236 LP –12243. https://doi.org/10.1523/JNEUROSCI.2378-09.2009

Mobbs, D., Petrovic, P., Marchant, J. L., Hassabis, D., Weiskopf, N., Seymour, B., Dolan, R. J., & Frith, C. D. (2007). When Fear Is Near: Threat Imminence Elicits Prefrontal-Periaqueductal Gray Shifts in Humans [doi: 10.1126/science.1144298]. *Science*, *317*(5841), 1079–1083. https://doi.org/10.1126/science.1144298

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

Morey, R. (2019). Why you shouldn't say "this study is underpowered". *Towards Data Science*.

Morey, R., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. https://doi.org/https://doi.org/10.5281/zenodo.838685

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morey, R. D., & Rouder, J. N. (2024). *Bayesfactor: Computation of bayes factors for common designs* [R package version 0.9.12-4.7]. https://CRAN.R-project.org/package=BayesFactor

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, *75*(1), 73–79. https://doi.org/10.1016/j.neuron.2012.05.021

Mounjid, O., & Lehalle, C. (2023). Improving reinforcement learning algorithms: Towards optimal learning rate policies. *Mathematical Finance*, *34*(2), 588–621. https://doi.org/10.1111/mafi.12378

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and cognition*, *25*, 67–76.

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.

Mulder, J., Hoijtink, H., & De Leeuw, C. (2012). Biems: A fortran 90 program for calculating bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*, 1–39.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of statistical planning and inference*, *140*(4), 887–906.

Mulder, J., Wagenmakers, E.-J., & Marsman, M. (2022). A Generalization of the Savage–Dickey Density Ratio for Testing Equality and Order Constrained Hypotheses. *The American Statistician*, *76*(2), 102–109. https://doi.org/10.1080/00031305.2020.1799861

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink-Bosman, M., Menke, J., van Aert, R., Fox, J.-P., et al. (2021). Bf-pack: Flexible bayes factor testing of scientific theories in r. *Journal of Statistical Software*, *100*, 1–63.

Mumford, J. A., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of regressors in fmri models. *PloS one*, *10*(4), e0126255.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. (2012). An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron*, *75*(6), 1114–1121. https://doi.org/10.1016/j.neuron.2012.07.023

Nitschke, J. P., Forbes, P. A. G., & Lamm, C. (2022). Does Stress Make Us More—or Less—Prosocial? A Systematic Review and Meta-Analysis of the Effects of Acute Stress on Prosocial Behaviours Using Economic Games. *Neuroscience & Biobehavioral Reviews*, 104905.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Norbury, A., Robbins, T. W., & Seymour, B. (2018). Value generalization in human avoidance learning. *eLife*, *7*, 1–30. https://doi.org/10.7554/eLife.34779

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: Ii. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on psychological science*, *7*(6), 615–31. https://doi.org/10.1177/1745691612459058

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669), 452–454. https://doi.org/10.1126/science.1094285

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53. https://doi.org/10.1196/annals.1390.022

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118.

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, *77*(3), 247–265. https://doi.org/10.1016/j.biopsycho.2007.11.006

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212.

Olsson, A., McMahon, K., Papenberg, G., Zaki, J., Bolger, N., & Ochsner, K. N. (2016). Vicarious Fear Learning Depends on Empathic Appraisals and Trait Empathy. *Psychological science*, *27*(1), 25–33. https://doi.org/10.1177/0956797615604124

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*(1), 3–11. https://doi.org/10.1093/scan/nsm005

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Padilla-Walker, L. M., Nelson, L. J., Carroll, J. S., & Jensen, A. C. (2010). More than a just a game: Video game and internet use during emerging adulthood. *Journal of Youth and Adolescence*, *39*(2), 103–113. https://doi.org/10.1007/s10964-008-9390-8

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature neuroscience*, *5*(2), 97–98. https://doi.org/10.1038/nn802

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z. L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48. https://doi.org/10.1016/j.jmp.2018.03.003

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

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

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, *25*(1), 127–154. https://doi.org/10.1007/s10851-006-6228-4

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Petrovic, P., Dietrich, T., Fransson, P., Andersson, J., Carlsson, K., & Ingvar, M. (2005). Placebo in Emotional Processing— Induced Expectations of Anxiety Relief Activate a Generalized Modulatory Network. *Neuron*, *46*(6), 957–969. https://doi.org/10.1016/j.neuron.2005.05.023

Pezzulo, G., Parr, T., & Friston, K. (2024). Active inference as a theory of sentient behavior. *Biological Psychology*, *186*, 108741.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*.

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology*, *15*(6), e1007043. https://doi.org/10.1371/journal.pcbi.1007043

Piray, P., Ly, V., Roelofs, K., Cools, R., & Toni, I. (2019). Emotionally Aversive Cues Suppress Neural Systems Underlying Optimal Learning in Socially Anxious

Individuals. *The Journal of Neuroscience*, *39*(8), 1445–1456. https://doi.org/10.1523/JNEUROSCI.1394-18.2018

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M., & Rawlins, J. N. (1999). Dissociating Pain from Its Anticipation in the Human Brain. *Science*, *284*(5422), 1979–1981. https://doi.org/10.1126/science.284.5422.1979

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32*(2), 197–205. https://doi.org/10.1016/j.neubiorev.2007.07.007

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Prescott, A. T., Sargent, J. D., & Hull, J. G. (2018). Metaanalysis of the relationship between violent video game play and physical aggression over time. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9882–9888. https://doi.org/10.1073/pnas.1611617114

Pruim, R. H., Mennes, M., Buitelaar, J. K., & Beckmann, C. F. (2015). Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *NeuroImage*, *112*, 278–287. https://doi.org/10.1016/j.neuroimage.2015.02.063

Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). Ica-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage*, *112*, 267–277. https://doi.org/10.1016/j.neuroimage.2015.02.064

Quesque, F., & Brass, M. (2019). The Role of the Temporoparietal Junction in Self-Other Distinction. *Brain Topography*, *32*, 943–955. https://doi.org/10.1007/s10548-019-00737-5

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*(1), 84–95. https://doi.org/10.1080/00223891.2010.528484

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, *72*, 33–37. https://doi.org/10.1016/j.jmp.2015.08.002

Robert, C. P., Casella, G., & Casella, G. (1999). *Monte carlo statistical methods* (Vol. 2). Springer.

Romero, F., & Sprenger, J. (2021). Scientific self-correction: The bayesian way. *Synthese*, *198*(Suppl 23), 5803–5823.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, *21*, 301–308.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for anova designs. *Journal of mathematical psychology*, *56*(5), 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., & Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature Neuroscience*, *17*(11), 1607–1612. https://doi.org/10.1038/nn.3832

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological bulletin*, *57*(5), 416.

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

Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., & Gläscher, J. (2020). Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, *146*, 107488.

Rütgen, M., Seidel, E. M., Riečanský, I., & Lamm, C. (2015). Reduction of empathy for pain by placebo analgesia suggests functional equivalence of empathy and first-hand emotion experience. *Journal of Neuroscience*, *35*(23), 8938–8947. https://doi.org/10.1523/JNEUROSCI.3936-14.2015

Rütgen, M., Seidel, E. M., Silani, G., Riečanský, I., Hummer, A., Windischberger, C., Petrovic, P., & Lamm, C. (2015). Placebo analgesia and its opioidergic regulation suggest that empathy for pain is grounded in self pain. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), E5638–E5646. https://doi.org/10.1073/pnas.1511269112

Rütgen, M., Wirth, E.-M., Riečanský, I., Hummer, A., Windischberger, C., Petrovic, P., Silani, G., & Lamm, C. (2021). Beyond Sharing Unpleasant Affect-Evidence for Pain-Specific Opioidergic Modulation of Empathy for Pain. *Cerebral cortex (New York, N.Y. : 1991)*, *31*(6), 2773–2786. https://doi.org/10.1093/cercor/bhaa385

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on bayesian hypothesis tests. *Psychonomic bulletin & review*, *21*, 283–300.

Sarafoglou, A., Haaf, J. M., Ly, A., Gronau, Q. F., Wagenmakers, E.-J., & Marsman, M. (2023). Evaluating multinomial order restrictions with bridge sampling. *Psychological Methods*, *28*(2), 322–338. https://doi.org/10.1037/met0000411

Schaaf, J. V., Weidinger, L., Molleman, L., & van den Bos, W. (2023). Test–retest reliability of reinforcement learning parameters. *Behavior Research Methods*, *56*(5), 4582–4599. https://doi.org/10.3758/s13428-023-02203-4

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of

mind: A review using probabilistic atlases from different imaging modalities. *Human brain mapping*, *38*(9), 4788–4805. https://doi.org/10.1002/hbm.23675

Seid-Fatemi, a., & Tobler, P. N. (2015). Efficient learning mechanisms hold in the social domain and are implemented in the medial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *10*(5), 735–743. https://doi.org/10.1093/scan/nsu130

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*(6992), 664–667. https://doi.org/10.1038/nature02581

Shao, R., & Wang, Y. (2019). The Relation of Violent Video Games to Adolescent Aggression: An Examination of Moderated Mediation Effect. *Frontiers in Psychology*, *0*, 384. https://doi.org/10.3389/FPSYG.2019.00384

Shields, G. S., Sazma, M. A., & Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience and Biobehavioral Reviews*, *68*, 651–668. https://doi.org/10.1016/j.neubiorev.2016.06.038

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09962-9

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *Journal of Neuroscience*, *33*(39), 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. *Science*, *303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*(10), 635–644. https://doi.org/10.1038/s41583-019-0180-y

Stan Development Team. (2024). *Stan modeling language users guide and reference manual*. Version 2.35. https://mc-stan.org

Statista Research Department. (2022). U.S. daily time spent playing games 2022, by age. [[Online; accessed 2023-08-29]]. https://www.statista.com/statistics/789835/average-daily-time-playing-games-us-by-age/

Staude-Müller, F., Bliesener, T., & Luthman, S. (2008). Hostile and hardened? An experimental study on (De-)Sensitization to violence and suffering through playing video games. *Swiss Journal of Psychology*, *67*(1), 41–50. https://doi.org/10.1024/1421-0185.67.1.41

Stefan, A. M., Katsimpokis, D., Gronau, Q. F., & Wagenmakers, E.-J. (2022). Expert agreement in prior elicitation and its effects on bayesian inference. *Psychonomic Bulletin & Review*, *29*(5), 1776–1794.

Steingroever, H., Wetzels, R., & Wagenmakers, E. J. (2013). Validating the PVL-Delta model for the Iowa gambling task. *Frontiers in Psychology*, *4*, 1–17. https://doi.org/10.3389/fpsyg.2013.00898

Steingroever, H., Wetzels, R., & Wagenmakers, E. J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *1*(3), 161–183. https://doi.org/10.1037/dec0000005

Steyvers, M., Lee, M. D., & Wagenmakers, E. J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179. https://doi.org/10.1016/j.jmp.2008.11.002

Stijovic, A., Forbes, P. A. G., Tomova, L., Skoluda, N., Feneberg, A. C., Piperno, G., Pronizius, E., Nater, U. M., Lamm, C., & Silani, G. (2023). Homeostatic Regulation of Energetic Arousal During Acute Social Isolation: Evidence From the Lab and the Field. *Psychological science*, *34*(5), 537–551. https://doi.org/10.1177/09567976231156413

Su, Z., Garvert, M. M., Zhang, L., Vogel, T. A., Cutler, J., Husain, M., Manohar, S. G., & Lockwood, P. L. (2025). Dorsomedial and ventromedial prefrontal cortex lesions differentially impact social influence and temporal discounting. *PLoS Biology*, *23*(4), e3003079.

Sun, S., & Yu, R. (2014). The feedback related negativity encodes both social rejection and explicit social expectancy violation. *Frontiers in Human Neuroscience*, *8*, 1–9. https://doi.org/10.3389/fnhum.2014.00556

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press Cambridge.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2012). Learning to Simulate Others' Decisions. *Neuron*, *74*(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

Swart, J. C., Froböse, M. I., Cook, J. L., Geurts, D. E., Frank, M. J., Cools, R., & den Ouden, H. E. (2017). Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *eLife*, *6*, 1–36. https://doi.org/10.7554/eLife.22169

Szycik, G. R., Mohammadi, B., Hake, M., Kneer, J., Samii, A., Münte, T. F., & te Wildt, B. T. (2017). Excessive users of violent video games do not show emotional desensitization: An fMRI study. *Brain Imaging and Behavior*, *11*(3), 736–743. https://doi.org/10.1007/s11682-016-9549-y

Szycik, G. R., Mohammadi, B., Münte, T. F., & te Wildt, B. T. (2017). Lack of evidence that neural empathic responses are blunted in excessive users of violent video games: An fMRI study. *Frontiers in Psychology*, *8*, 174. https://doi.org/10.3389/fpsyg.2017.00174

Tabor, A., & Burr, C. (2019). Bayesian Learning Models of Pain: A Call to Action. *Current Opinion in Behavioral Sciences*, *26*, 54–61. https://doi.org/10.1016/j.cobeha.2018.10.006

Thorndike, E. L. (1927). The Law of Effect. *The American Journal of Psychology*, *39*(1/4), 212. https://doi.org/10.2307/1415413

Tomova, L., Majdandžic, J., Hummer, A., Windischberger, C., Heinrichs, M., & Lamm, C. (2017). Increased neural responses to empathy for pain might explain how acute stress increases prosociality. *Social cognitive and affective neuroscience*, *12*(3), 401–408. https://doi.org/10.1093/scan/nsw146

Tomova, L., Saxe, R., Klöbl, M., Lanzenberger, R., & Lamm, C. (2020). Acute stress alters neural patterns of value representation for others. *NeuroImage*, *209*, 116497. https://doi.org/10.1016/j.neuroimage.2019.116497

Tortolero, S. R., Peskin, M. F., Baumler, E. R., Cuccaro, P. M., Elliott, M. N., Davies, S. L., Lewis, T. H., Banspach, S. W., Kanouse, D. E., & Schuster, M. A. (2014). Daily violent video game playing and depression in preadolescent youth. *Cyberpsychology, Behavior, and Social Networking*, *17*(9), 609–615. https://doi.org/10.1089/cyber.2014.0091

Tran, D., Hoffman, M. D., Moore, D., Suter, C., Vasudevan, S., Radul, A., Johnson, M., & Saurous, R. A. (2018). Simple, distributed, and accelerated probabilistic programming. *Neural Information Processing Systems*.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/NIMG.2001.0978

Valton, V., Wise, T., & Robinson, O. J. (2020). The importance of group specification in computational modelling of behaviour. *PsyArXiv*. https://doi.org/10.31234/osf.io/p7n3h

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09161-6

van de Schoot, R., & Strohmeier, D. (2011). Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioral Development*, *35*(2), 180–190. https://doi.org/10.1177/0165025410397432

Vandendriessche, H., Demmou, A., Bavard, S., Yadak, J., Lemogne, C., Mauras, T., & Palminteri, S. (2022). Contextual influence of reinforcement learning performance of depression: Evidence for a negativity bias? *Psychological Medicine*, *53*(10), 4696–4706. https://doi.org/10.1017/s0033291722001593

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. https://doi.org/10.1016/J.JMP.2010.07.003

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1–20. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., & Ojanen, J. (2012). A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*.

Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). Stop-IT: Windows executable software for the stop-signal paradigm. *Behavior Research Methods*, *40*(2), 479–483. https://doi.org/10.3758/BRM.40.2.479

Vlaeyen, J. W. (2015). Learning to predict and control harmful events. *PAIN*, *156*, S86–S93. https://doi.org/10.1097/j.pain.0000000000000107

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., et al. (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*, 35–57.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, *47*, 913–917.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2012). A Default Bayesian Hypothesis Test for ANOVA Designs. *The American Statistician*, *66*(2), 104–111. https://doi.org/10.1080/00031305.2012.695956

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wiech, K., & Tracey, I. (2013). Pain, decisions, and actions: A motivational perspective. *Frontiers in Neuroscience*, *7*, 46. https://doi.org/10.3389/fnins.2013.00046

Wiecki, T., Sofer, I., & Frank, M. J. (2013). Hddm: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 1–10. https://doi.org/10.3389/fninf.2013.00014

Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife*, *6*, e28098. https://doi.org/10.7554/eLife.28098

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, 1–35. https://doi.org/10.7554/eLife.49547

Wilson, R. C., & Niv, Y. (2015). Is Model Fitting Necessary for Model-Based fMRI? *PLoS Computational Biology*, *11*(6), 1–21. https://doi.org/10.1371/journal.pcbi.1004237

Yoon, L., Somerville, L. H., & Kim, H. (2018). Development of MPFC function mediates shifts in self-protective behavior provoked by social feedback. *Nature Communications*, *9*(1), 1–10. https://doi.org/10.1038/s41467-018-05553-2

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic bulletin & review*, *21*, 268–282.

Zaki, J., & Mitchell, J. P. (2013). Intuitive Prosociality. *Current Directions in Psychological Science*, *22*(6), 466–470. https://doi.org/10.1177/0963721413492764

Zhang, L., & Gläscher, J. P. (2019). A network supporting social influences in human decision-making. *bioRxiv*, 551614. https://doi.org/https://doi.org/10.1101/551614

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls, and suggestions. *PsyArXiv*, 10.31234/osf.io/uthw2. https://doi.org/10.31234/OSF.IO/UTHW2

Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences*, *109*(5), 1419–1424. https://doi.org/10.1073/pnas.1116783109

List of Figures

List of Tables

Zusammenfassung

In den letzten Jahren hat das Bayessche Denken in der Psychologie und den Neurowissenschaften eine bemerkenswerte Renaissance erfahren. In dieser kumulativen Dissertation untersuche ich, wie die Bayessche Inferenz unser Verständnis menschlichen Sozialverhaltens und seiner neuronalen Grundlagen vertiefen kann. Die Arbeit gliedert sich in zwei Teile, die den zweifachen Schwerpunkt meiner Forschung widerspiegeln.

Im ersten Teil befasse ich mich mit der Frage, wie Bayessche Methoden die Inferenz in klassischen experimentellen Designs verbessern können. In einer fMRT-Studie zu den Effekten gewalthaltiger Videospiele auf Empathie zeige ich, wie Bayessche Analysen Aussagen sowohl über das Vorhandensein als auch das Fehlen experimenteller Effekte ermöglichen. Anschließlich behandle ich in einem theoretischen Beitrag typische Missverständnisse über den Begriff der statistischer „Power“, und diskutiere Unterschiede zwischen bayesianischer und frequentistischer Inferenz.

Im zweiten Teil steht die Anwendung Bayesscher kognitiver Modelle – insbesondere von Reinforcement-Learning-Modellen – in den Sozialen Neurowissenschaften im Fokus. Ich präsentiere zunächst eine experimentelle fMRT-Studie zur neuronalen Basis prosozialen Lernens und Entscheidens. Aufbauend auf Erkenntnissen aus dieser empirischen Arbeit diskutiere ich anschließend typische Interpretationsfehler und „best practices“ beim Einsatz von Reinforcement-Learning-Modellen. Abschließend entwickle ich eine effiziente Methode zur Bayesschen Testung restringierter Hypothesen, und zeige exemplarisch, wie sich damit Theorien über menschliches Lernverhalten testen lassen.

Diese fünf Beiträge veranschaulichen, wie Bayessche Modellierung als pragmatisches Werkzeug zur Datenanalyse eingesetzt werden kann, aber auch wie Bayessches Denken unser Verständnis für den wissenschaftlichen Prozess an sich vertiefen kann. Ziel der Arbeit ist es, empirische Neurowissenschaft und statistische Theorie zu verbinden und das Potenzial eines vielfältigen methodischen Instrumentariums aufzuzeigen.

List of Publications

Published in Peer-Reviewed Journals

* indicates shared first authorship

- **Lengersdorff, L. L.** & Lamm, C. (2025). With Low Power comes Low Credibility? Towards a Principled Critique of Results from Underpowered Tests. *Advances in Methods and Practices in Psychological Science*, *8*(1). https://doi.org/10.1177/25152459241296397

- Steininger, M. O., White, M. P., **Lengersdorff, L. L.**, Zhang, L., Smalley, A. J., Kühn, S., & Lamm, C. (2025). Nature exposure induces analgesic effects by acting on nociception-related neural processing. *Nature Communications*, *16*(1), 2037. https://doi.org/10.1038/s41467-025-56870-2

- Boch, M., Karl, S., Wagner, I. C., **Lengersdorff, L. L.**, Huber, L., & Lamm, C. (2024). Action observation reveals a network with divergent temporal and parietal cortex engagement in dogs compared with humans. *Imaging Neuroscience, 2*, 1-29. https://doi.org/10.1162/imag_a_00385

- **Lengersdorff, L. L.**, Wagner, I. C., Mittmann, G., Sastre-Yagüe, D., Lüttig, A., Olsson, A., Petrovic, P., & Lamm, C. (2023). Neuroimaging and behavioral evidence that violent video games exert no negative effect on human empathy for pain and emotional reactivity to violence. *eLife*, *12*, e84951. https://doi.org/10.7554/eLife.84951

- Müllner-Huber, A., Anton-Boicuk, L., Pronizius, E., **Lengersdorff, L. L.** , Olsson, A., & Lamm, C. (2022). The causal role of affect sharing in driving vicarious fear learning. *Plos one, 17*(11), e0277793. https://doi.org/10.1371/journal.pone.0277793 https://doi.org/10.1371/journal.pone.0277793

- Stefan, A. M., **Lengersdorff, L. L.**, & Wagenmakers, E. J. (2022). A two-stage Bayesian sequential assessment of exploratory hypotheses. *Collabra: Psychology, 8*(1), 40350. https://doi.org/10.1525/collabra.40350

- Hartmann, H.*, **Lengersdorff, L. L.***, Hitz, H. H., Stepnicka, P., & Silani, G. (2022). Emotional ego-and altercentric biases in high-functioning autism spectrum disorder: Behavioral and neurophysiological evidence. *Frontiers in Psychiatry, 13*, 813969. https://doi.org/10.3389/fpsyt.2022.813969

- **Lengersdorff, L. L.**, Wagner, I. C., Lockwood, P. L., & Lamm, C. (2020). When implicit prosociality trumps selfishness: the neural valuation system underpins more optimal choices when learning to avoid harm to others than to oneself. *The Journal of Neuroscience*, *40*(38), 7286–7299. https://doi.org/10.1523/JNEUROSCI.0842-20.2020

- Zhang, L.*, **Lengersdorff, L. 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

– Riečanský, I.*, **Lengersdorff, L. L.***, Pfabigan, D. M., & Lamm, C. (2020). Increasing self-other bodily overlap increases sensorimotor resonance to others' pain. *Cognitive, Affective, & Behavioral Neuroscience, 20*, 19-33. https://doi.org/10.3758/s13415-019-00724-0

Preprinted or Submitted

– **Lengersdorff, L. L.**, & Marsman, M. (2025). Sampling in Constrained Space: Efficient Estimation of Model Evidence under Equality and Inequality Constraints. *PsyArXiv*. https://doi.org/10.31234/osf.io/a3ytg_v1

– Doell, K., **Lengersdorff, L. L.**, Rhoads, S., Todorova, B., Nitschke, J., Druckman, J., ... & Van Bavel, J. (2025). When Predicting Climate-Relevant Intervention Effectiveness, Academics Outperform the Public, but Not a Simple Heuristic. *PsyArXiv*. https://doi.org/10.31234/osf.io/gx7au_v1

– Todorova, B., Zhang, L., **Lengersdorff, L. L.**, Nitschke, J.P., Forbes, P., Doell, K.C., & Lamm, C. (submitted). Motivational asymmetries in discounting of effort and time in self-benefitting vs. pro-environmental decisions.

Technical Notes

This thesis was written in LaTeX using the online environment Overleaf:

https://www.overleaf.com/.

The thesis layout is based on the template for Masters and Doctoral theses available here:

https://www.latextemplates.com/template/masters-doctoral-thesis.

I used the reference extractor developed by Rintze M. Zelle (https://rintze.zelle.me/ref-extractor/) to convert reference information contained within Microsoft Word manuscript files into the BibTex format used in LaTeX.

To convert tables originally created in Microsoft Word to LaTeX, I used the table generator available here: https://www.tablesgenerator.com/