Machine Learning Methods For Modeling Gaze Allocation In Simple Choice Behavior And Functional Neuroimaging Data On The Level Of The Individual

Vorgelegt von M. Sc.
Armin W. Thomas
ORCID: 0000-0002-9947-5705

an der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Klaus Obermayer
Gutachter:	Prof. Dr. Klaus-Robert Müller
	Prof. Dr. Hauke R. Heekeren
	Prof. Dr. Russell A. Poldrack

Tag der wissenschaftlichen Aussprache: 03. November 2020

Berlin 2020

Abstract

The work presented in this thesis uses tools from machine learning, statistics, and computation to build analysis methods for capturing individual differences related to two research questions from cognitive neuroscience.

The *first chapter* of this thesis investigates the role of looking behavior in simple choices of individuals (e.g., deciding whether to eat an apple or a banana for breakfast). To this end, we introduce a decision model that allows us to study the association of looking behavior and choice on the level of the individual and in varying choice set sizes. By the use of this model, we provide empirical evidence that gaze allocation has an active influence on the decision process of the individual in choice situations with two to 36 choice alternatives, such that individuals are generally more likely to choose the alternative that they have looked at longer. We further demonstrate that the strength of this association varies between individuals and that accounting for this variability is necessary to accurately explain and predict individuals' choice behavior.

The *second chapter* of this thesis investigates the analysis of whole-brain functional Magnetic Resonance Imaging (fMRI) data through deep learning models. Over recent years, deep learning models have had strong empirical success due to their ability to learn complex behaviors and associate a target signal with variable patterns in high-dimensional and noisy datasets. This makes deep learning models well-suited for the analysis of fMRI data, where the mapping between cognitive state (e.g., deciding whether to accept or reject a risk) and brain activity can be widely spatially distributed and vary between individuals. To investigate the ability of deep learning models to capture this variability, we introduce and evaluate a framework for the analysis of fMRI data. This framework first uses a deep learning model to decode a cognitive state from a whole-brain fMRI volume. Subsequently, it relates the decoded cognitive state and fMRI data by the application of the layer-wise relevance propagation technique. We demonstrate that this framework can accurately decode a set of cognitive states from whole-brain fMRI data and identify a biologically plausible association between the decoded cognitive states and fMRI data. Importantly, it can do so on different levels of data granularity, from the level of the group down to the level of the individual and single time point. Lastly, we demonstrate that the performance of deep learning models in conventional fMRI datasets can be improved through transfer learning, by pre-training these models on large, publicly available fMRI datasets.

Zusammenfassung

Ziel dieser Dissertation ist es, mithilfe von Methoden des maschinellen Lernens, der Statistik und Informatik, Analysemethoden zu entwickeln, die es erlauben individuelle Unterschiede in zwei Forschungsfragen der kognitiven Neurowissenschaft zu erfassen.

Das *erste Kapitel* dieser Dissertation widmet sich der Rolle von Blickbewegungen in einfachen Entscheidungsprozessen des Individuums (z.B., die Entscheidung eine Banane oder einen Apfel zum Frühstück zu essen). Um diese Frage zu untersuchen, führen wir ein Entscheidungsmodell ein, das es erlaubt die Beziehung von Blickbewegungen und Entscheidungsverhalten auf der Ebene des Individuums zu untersuchen und das auf Entscheidungssituationen mit vielen Alternativen anwendbar ist. Mit diesem Modell zeigen wir, dass Blickbewegungen einen aktiven Einfluss auf das Entscheidungsverhalten des Individuums haben, so dass Individuen sich eher für diejenigen Alternativen entscheiden, die sie länger angesehen haben. Dieser Zusammenhang ist in Entscheidungssituationen mit wenigen und vielen Alternativen aufzufinden und variiert in seiner Stärke zwischen Individuen. Wir zeigen zudem, dass es notwendig ist diese individuellen Unterschiede zu berücksichtigen, um Entscheidungsverhalten akkurat erklären und vorhersagen zu können.

Das *zweite Kapitel* dieser Dissertation widmet sich der Analyse von funktionellen Magnetresonanztomographie (fMRT) Daten durch tiefe neuronale Netzwerke. Tiefe neuronale Netzwerke haben derzeit grossen empirischen Erfolg, aufgrund ihrer Fähigkeit komplexe Verhaltensweisen zu lernen und Signale mit variablen Mustern in hochdimensionalen und verrauschten Daten in Verbindung zu bringen. Aufgrund dieser Fähigkeit sind sie vielversprechend für die Analyse von fMRT Daten, da der Zusammenhang zwischen einem kognitiven Zustand (z.B., bei der Entscheidung eine risikoreiche Wette einzugehen oder abzulehnen) und der zugrundeliegenden Hirnaktivität räumlich weit verbreitet und zwischen Individuen variabel sein kann. Um die Fähigkeit tiefer neuronaler Netzwerke zu untersuchen diese Variabilität zu erfassen, führen wir eine Analysemethode für fMRT Daten ein. Diese Methode nutzt zuerst ein tiefes neuronales Netzwerk, um einen kognitiven Zustand aus fMRT Daten auszulesen. Im Anschluss, bringt sie diese Entscheidung mit den fMRT Daten in Verbindung, durch die Anwendung von layer-wise relevance propagation. Wir zeigen, dass diese Analysemethode akkurat kognitive Zustände aus fMRT Daten erkennt und biologisch bedeutungsvolle Zusammenhänge zwischen diesen Zuständen und der zugrundeliegenden Hirnaktivität identifiziert. Wir zeigen zudem, dass die Leistung von tiefen neuronalen Netzwerken in konventionellen fMRT Datensätzen verbessert werden kann, indem man sie zuvor auf grossen, öffentlichen fMRT Datensätzen trainiert.

Acknowledgments

First and foremost, I would like to thank Klaus-Robert Müller and Hauke R. Heekeren for being exceptional advisors and for providing me with the opportunity to pursue the research presented in this thesis. Klaus and Hauke both helped me to stay enthusiastic, persistent, and focused in moments in which nothing added up. They are invaluable sources of motivation and inspiration, for which I am very thankful. I would also like to thank Felix Molter, Wojciech Samek, Ian Krajbich, and Peter N. C. Mohr, who collaborated with me on the research presented in this thesis. Without their contributions, this work would have not been possible. I would further like to thank Antonio Rangel who inspired me to pursue a doctorate through his devotion to research. I would also like to thank my brother, my parents, and my friends for their continuing support and for being role models in my life. I further thank my colleagues at Freie and Technische Universität Berlin for many inspiring discussions, including: Sebastian Lapuschkin, Rasmus Bruckner, Vignesh Srinivasan, and Maximilan Kohlbrenner. My thesis was supported by the Bernstein Center for Computational Neuroscience in Berlin as well as the Max Planck School of Cognition.

Contents

Chapter 0

General introduction

The range of human behavior and cognition is highly diverse. We differ, for example, in our preferences for music and food, our political beliefs, and our willingness to take risks. A consensus among contemporary scientists is that our thoughts, feelings, actions, and thereby our behavioral and cognitive differences, are linked to the neurobiological processes of our central and peripheral nervous system and the interaction of these processes with the influences of our environment.

Understanding the association between the neurobiological processes of the brain and our individual cognitive states (e.g., when deciding to accept or reject a risky gamble) is one of the fundamental goals of cognitive neuroscience research. However, as neuroscience data is generally scarce, noisy, and high-dimensional, studying these associations on the level of the individual is challenging. Further, building and evaluating models on the level of the individual can be computationally and theoretically intractable for many research questions.

In a first step towards understanding the association between cognitive states and brain activity, researchers have therefore often focused on the group-level. Aggregating over the differences between individuals generally simplifies the hypothesis to be tested, while increasing the statistical power of the test. In addition, group-level findings are important building blocks of the general understanding of the mapping between cognitive states and brain activity.

Nevertheless, group-level models are not always appropriate (e.g., to understand individual differences in risk taking behavior) and can be misleading about the individual in the group (a group-level model on risk taking would falsely predict some individuals to be too risk averse and others to be too risk seeking). Further, studying individual differences in the association between cognitive states, behavior, and brain activity can help us build more comprehensive models of these associations that better account for their dynamics.

The work presented in this thesis uses tools from machine learning, statistics, and computation to build analysis frameworks for capturing such individual differences related to two research questions from cognitive neuroscience.

The **first chapter** of this thesis investigates the role of looking behavior (i.e., gaze allocation) in simple decision processes of the individual. Imagine deciding whether to eat an apple or an orange from a fruit basked in front of you or choosing which movie to watch on Netflix. In these simple choice situations, our eyes search through the set of available alternatives, and we might go back and forth between several alternatives to compare them. Recent empirical work has indicated that the duration and sequence

in which we look through the set of alternatives influences our choices, with generally higher choice probabilities for those alternatives that we look at longer. This finding motivated the development of a set of computational models that integrate eye movement data into the choice process and formalize the empirically observed association between gaze and choice. These models are based on classical evidence accumulation models, but make the additional assumption that the momentary rate of evidence accumulation depends on the eye movements of the decision-maker.

While these models have been very successful in explaining simple choice behavior (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011), they are often computationally and analytically intractable. This has limited their application to group-level data and choice sets with only few alternatives. It is therefore unclear whether an association between gaze allocation and choice behavior is actually present on the level of the individual and whether it generalizes to choice situations involving many alternatives. To investigate these questions, we introduce a decision model (the gaze-weighted linear accumulator model, or GLAM), which allows us to study the association of gaze and choice on the level of the individual and in choice situations with many alternatives. With the GLAM, we study five choice datasets, spanning 167 individuals, six choice set sizes (2, 3, 9, 16, 25, and 36 alternatives), and two choice domains (value-based and perceptual choice). We find that I) an association between gaze allocation and choice behavior is present on the level of the individual across all choice set sizes and domains; II) the strength of the association between gaze and choice is variable between individuals; and III) this inter-individual variability needs to be taken into account to accurately explain and capture choice behavior.

The **second chapter** of this thesis investigates the application of deep learning (DL; Goodfellow et al., 2016, LeCun et al., 2015) models to the analysis of whole-brain functional Magnetic Resonance Imaging (fMRI) data. Due to the high spatial and low temporal resolution of fMRI, these datasets are typically very high-dimensional (with a few hundred thousand dimensions (or voxels) per fMRI volume), while containing comparably few volumes for each individual in the data. For this reason, many conventional approaches for the analysis of fMRI data include restricting assumptions by analyzing the data of individual voxels or groups of voxels independent of one another, using simple linear mappings between the cognitive states and brain activity, or solely focusing on the level of the group. However, functional neuroimaging research has recently experienced an increase in the availability of large and public fMRI datasets (e.g., Casey et al., 2018, Sudlow et al., 2015, Van Essen et al., 2013). The availability of these datasets has paved the way to explore individual differences in the association between cognitive states and brain activity, and to apply advanced computational methods that can better account for the non-linear, temporo-spatial variability of whole-brain activity within and between individuals. One of these methods, with strong empirical success in many research and industry applications, is DL. DL models are able to learn complex behaviors and associate a target signal with variable patterns in high-dimensional and noisy datasets. This makes DL methods especially promising for the analysis of whole-brain functional neuroimaging data, where the mapping between cognitive states and brain activity can be widely spatially distributed and variable between individuals.

In spite of this promise, two major challenges have so far prevented broad DL usage in functional neuroimaging research: First, even if a DL model is able to successfully decode a set of cognitive states from fMRI data, the nature of the learned relationship between fMRI data and decoded cognitive state can not be directly understood, as it is disguised by the highly nonlinear nature of DL models. Second, in the high-dimension, low-sample size setting of typical fMRI datasets, DL methods (as well as conventional machine learning approaches) are prone to overfitting.

The goal of the work presented in this thesis is to investigate two solutions to these challenge, which utilize methods of explainable artificial intelligence and transfer learning. First, we introduce the DeepLight framework, which uses a DL model to identify a set of cognitive states from whole-brain fMRI data and, subsequently, relates the decoded cognitive state and brain activity by interpreting the individual decoding decisions with the layer-wise relevance propagation technique (LRP; Bach et al., 2015, Montavon et al., 2017, Samek et al., 2020). In an empirical comparison of DeepLight to three conventional fMRI analysis approaches, we demonstrate that DeepLight can accurately decode a set of cognitive states from whole-brain fMRI data and identify a biologically plausible association between these cognitive states and the fMRI data; on the level of the group, individual, and single time points. We further investigate whether transfer learning can improve the decoding performance of DL models in conventional fMRI datasets. To this end, we train DeepLight on a large, public fMRI dataset of the Human Connectome Project (HCP; Van Essen et al., 2013), spanning 400 individuals and six experiment tasks. We demonstrate that this pre-trained model generally learns quicker, achieves higher decoding accuracies, and requires less training data than a model variant that is not pre-trained, when both are applied to the fMRI data of the left-out seventh HCP experiment task. Interestingly, the pre-trained model at first does not exhibit the same advantages when applied to an fMRI dataset that is not part of the HCP. Nevertheless, when accounting for both datasets during pre-training, the pre-trained model again clearly outperforms a model variant that is not pre-trained at decoding the cognitive states from the independent fMRI dataset, demonstrating that transfer learning is, in general, beneficial for the application of DL models to conventional fMRI datasets.

0.1 Structure of the thesis

The two main chapters of this thesis have the same overall structure. Both provide an introduction to their research question, present the methodology and results of a set of empirical studies, and provide a conclusion and discussion of the empirical findings.

Chapter I: The role of gaze allocation in simple choice

Introduction In this section, we will provide an overview of existing empirical work on the role of gaze allocation in simple choice and introduce the attentional drift-diffusion model (aDDM; Krajbich et al., 2010, Krajbich and Rangel, 2011), which builds the conceptual basis for the gaze-weighted linear accumulator model (GLAM; see section 1.3.2) that we propose. Subsequently, we will highlight several practical limitations of the aDDM and outline the need to study individuals' choice behavior, as findings on the group-level can be misleading about the individuals in the group. We will further introduce other decision-making frameworks that have been proposed for choices from many alternatives.

Research questions and hypotheses In this section, we will specify a set of research questions and hypotheses that will be tested in this work.

Methodology In this section, we will introduce the five choice datasets, spanning 167 individuals, six choice set sizes (2, 3, 9, 16, 25, and 36 alternatives), and two choice domains (value-based and perceptual choice) that we analyzed in this chapter of the thesis. We will further provide the methodological details of the gaze-weighted linear accumulator model and a set of competing choice models for many-alternative choice situations.

Results In this section, we will present the results of two empirical studies that we performed. The first study investigates the association of gaze allocation and choice behavior on the level of the individual. The second study investigates this association in choice situations with many alternatives. We will further present an exploratory analysis of individuals' visual search behavior in choice situations with many alternatives.

Conclusion, Discussion Lastly, we will provide an overall conclusion of our empirical findings with respect to our research hypotheses. We will further provide a discussion of the empirical findings as well as the strengths and limitations of the GLAM .

Chapter II: The analysis of functional neuroimaging data through deep learning models

Introduction In this section, we will give an introduction to general challenges for the analysis of fMRI data and subsequently outline the premise and challenges for the application of DL methods to fMRI data. We will further introduce explanation techniques and transfer learning as solutions to these challenges.

Research questions and hypotheses In this section, we will specify a set of research questions and hypotheses that will be tested in this work.

Methodology In this section, we will introduce the two fMRI datasets that we analyzed in this chapter of the thesis, spanning more than 450 individuals and eight experiment tasks. We will further provide a detailed overview of the DeepLight framework and three conventional analysis approaches for fMRI data.

Results In this section, we will present the results of two empirical studies that we performed. The first study investigates the performance of DeepLight in accurately decoding a set of cognitive states from whole-brain fMRI data and identifying a biologically plausible association between these cognitive states and the underlying brain activity. The second study investigates whether transfer learning can be used to improve the application of DL models to conventional fMRI datasets. We will further provide a comparison of two different DeepLight architectures in our transfer learning analyses.

Conclusion, Discussion Lastly, we will provide an overall conclusion of our empirical findings with respect to our research hypotheses. We will further provide a discussion of the strengths and limitations of the DeepLight framework in comparison to conventional analysis approaches for fMRI data, and discuss the potential and challenges for transfer learning with fMRI datasets.

0.2 Main contributions

This thesis makes the following theoretical, empirical, and methodological contributions:

Contributions to research on the role of gaze allocation in simple choice

1. The gaze-weighted linear accumulator model (GLAM). We introduce a decision model (the gaze-weighted linear accumulator model GLAM; see section 1.3.2) which allows to study the association of gaze allocation and choice behavior on the level of the individual (solely by the use of choice, RT, gaze, and liking rating data) and generalizes to choice situations involving many alternatives.

2. GLAMbox. To make the GLAM accessible to a wide audience of researchers, we created a Python-based toolbox (GLAMbox; `https://glambox.readthedocs.io/en/latest/`) that is build on top of PyMC3 (Salvatier et al., 2016) and allows the easy application of the GLAM to experimental choice data (for details, see section 1.3.3).

3. A choice dataset with many alternatives. We collected a novel choice dataset, in which 49 individuals make simple value-based choices from sets of 9, 16, 25, and 36 snack food items. This dataset includes the gaze, choice, RT, and liking rating data for each of 50 trials per individual in each of the four choice set sizes. This dataset will be made publicly available at `https://github.com/athms/many-item-choice`.

4. Two empirical studies. In an analysis of five choice datasets, spanning 167 individuals, six choice set sizes (2, 3, 9, 16, 25, and 36 alternatives), and two choice domains (value-based and perceptual choice), we provide empirical evidence that I) an association of gaze allocation and choice behavior is present on the level of the individual, across all choice set sizes and domains; II) the strength of this association is variable between individuals; and III) it is necessary to account for this inter-individual variability to accurately explain and predict individuals' choice behavior. We further provide an exploratory analysis of individuals' visual search behavior in choice sets with many alternatives.

Contributions to the analysis of functional neuroimaging data through deep learning models

1. The DeepLight framework. We introduce the DeepLight framework, which utilizes a DL model to accurately decode a set of cognitive states from whole-brain fMRI data and subsequently relates the decoded cognitive state and brain activity by the use of the LRP technique (Bach et al., 2015, Montavon et al., 2017, Samek et al., 2020).

2. Two DeepLight architectures. We introduce and compare two DeepLight architectures and provide a specification of the LRP technique for each architecture.

3. A transfer learning framework for fMRI data. We introduce and evaluate a mechanism for transfer learning with fMRI datasets and compare two different fine-tuning approaches.

4. Two empirical studies. In two studies, spanning the fMRI data of more than 450 individuals and eight experiment tasks, we provide empirical evidence that I) the combination of DL methods and explanation techniques can be used to accurately decode a set of cognitive states from whole-brain fMRI data and identify a biologically plausible association between brain activity and cognitive state and II) transfer learning can generally improve the decoding performance of DL models in conventional fMRI datasets.

0.3 List of publications

Many findings of this thesis have been previously published in the following peer-reviewed journal publications, preprints, and conference proceedings.

- Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., & Mohr, P. N. (2019). Gaze bias differences capture individual choice behaviour. Nature human behaviour, 3(6), 625-635. doi.org/10.1038/s41562-019-0584-8
- Thomas, A. W., Molter, F., Heekeren, H. R., & Mohr, P. N. (2019). GLAMbox: A Python toolbox for investigating the association between gaze allocation and decision behaviour. PLOS ONE, 14(12):e0226428.

doi.org/10.1371/journal.pone.0226428

- Thomas, A. W., Heekeren, H. R., Müller, K. R., & Samek, W. (2019). Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. Frontiers in Neuroscience, 13:1321. doi.org/10.3389/fnins.2019.01321

- Thomas, A. W., Müller, K. R., & Samek, W. (2019). Deep Transfer Learning for Whole-Brain FMRI Analyses. In: Zhou L. et al. (eds) OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging. OR 2.0 2019, MLCN 2019. Lecture Notes in Computer Science, vol 11796. Springer, Cham. doi.org/10.1007/978-3-030-32695-1_7

- Thomas, A. W., Molter, F., & Krajbich, I. (2020, PsyArXiv preprint). Uncovering the Computational Mechanisms Underlying Many-Alternative Choice. doi.org/10.31234/osf.io/tk6qe

I would like to thank my co-authors for allowing me to use text and figures from these publications in this thesis.

Chapter 1

The role of gaze allocation in simple choice

This chapter covers the contributions from Thomas et al. (2019b) Thomas et al. (2020), and Molter et al. (2019). I would like to thank my co-authors for allowing me to use text and figures from these publications.

1.1 Introduction

1.1.1 Gaze allocation and simple choice behavior

In everyday life, we are constantly confronted with simple consumer choices, such as whether to have an apple or a banana for breakfast or which bottle of juice to buy at the supermarket. Traditional models describing this type of choice assume that people assign a utility (or value) to each available option and make utility-maximizing choices (Fehr and Rangel, 2011, Glimcher and Fehr, 2013, von Neumann and Morgenstern, 2007). Notably, choices are assumed to be based solely on the attributes of the option, and are therefore independent of information search processes during decision formation (Luce and Raiffa, 1957). This assumption has recently been challenged by a variety of empirical findings showing that the allocation of gaze during the decision-making process also plays a substantial role, as a longer gaze towards one alternative is regularly associated with a higher choice probability for that alternative (e.g., Armel et al., 2008, Cavanagh et al., 2014, Fiedler and Glöckner, 2012, Folke et al., 2016, Glöckner and Herbold, 2011, Konovalov and Krajbich, 2016, Krajbich et al., 2010, 2015, 2012, Krajbich and Rangel, 2011, Pärnamets et al., 2015, Shimojo et al., 2003, Stewart et al., 2016a,b, Vaidya and Fellows, 2015). Similarly, stimulus salience and the external manipulation of gaze have been shown to influence decision behavior and choice probabilities (e.g., Armel et al., 2008, Milosavljevic et al., 2012, Pärnamets et al., 2015, Shimojo et al., 2003, Towal et al., 2013, Tsetsos et al., 2012). Similar effects have recently also been demonstrated in perceptual decision-making, where participants judge perceptual attributes of stimuli based on the available sensory information (for example, the orientation of line segments; Tavares et al., 2017).

These findings led to the development of computational models, such as the attentional Drift Diffusion Model (aDDM; for further details, see section 1.1.2 and Krajbich et al., 2010, Krajbich and Rangel, 2011), that integrate eye movement data into the decision process and thereby formalize the empirically observed association between gaze and choice. These models are based on classical evidence accumulation models (Ratcliff,

1978, Ratcliff et al., 2016), but make the additional assumption that the momentary rate of evidence accumulation depends on the allocation of gaze of the decision-maker. Evidence accumulation for an option is assumed to be discounted while another item is looked at. By accounting for this gaze bias, these models provide a precise account of many aspects of simple choice behavior at the group level and in choice situations with two to four alternatives (e.g., Cavanagh et al., 2014, Fisher, 2017, Gluth et al., 2018, Krajbich et al., 2010, 2012, Krajbich and Rangel, 2011, Towal et al., 2013).

1.1.2 The attentional drift-diffusion model

The attentional Drift-Diffusion Model (aDDM), which was first proposed by Krajbich and colleagues (Krajbich et al., 2010), is one of the most prominent models that integrate eye movement data into the decision process. The aDDM extends the classical binary Drift-Diffusion Model (DDM; Ratcliff, 1978, Ratcliff and McKoon, 2007) by the influence of gaze allocation onto the decision process.

Specifically, the aDDM includes an evidence accumulator E_i for each alternative i in a choice set. Each of these evidence accumulators can be in two states: When alternative i is looked at, E_i is defined as:

$$E_i(t) = E_i(t-1) + v \cdot r_i + \mathcal{N}(0, \sigma^2) \text{ with } E_i(0) = 0 \tag{1.1}$$

Here, r_i describes the value of alternative i, while v represents a general velocity parameter that determines the overall speed of evidence accumulation, and σ the standard deviation of zero-centered normal-distributed accumulation noise, which is added at each time step t (with $E_i(0) = 0$). For all other alternatives j, which are momentarily not looked at, evidence is accumulated at a discounted rate:

$$E_i(t) = E_i(t-1) + v \cdot \theta \cdot r_i + \mathcal{N}(0, \sigma^2) \text{ with } E_i(0) = 0 \tag{1.2}$$

The strength of the discounting of the accumulated evidence (the gaze bias) is determined by θ ($0 < \theta < 1$).

The aDDM assumes that choices are driven by a relative decision value RDV_i for each choice alternative i. The RDV is defined as the difference between the alternative's accumulated evidence E_i and the maximum accumulated evidence of all other alternatives j:

$$RDV_i(t) = E_i(t) - max_{j \neq i} E_j(t) \tag{1.3}$$

As soon as any RDV reaches a common decision boundary b (which is generally set to 1; Krajbich et al., 2010, Krajbich and Rangel, 2011), a choice is made for the respective alternative.

Note that for choices between only two alternatives, the aDDM can be reduced to a single RDV that diffuses between two symmetric decision boundaries ($+1$ and -1, each representing one of the two alternatives):

$$RDV(t) = E_{+1}(t) - E_{-1}(t) \tag{1.4}$$

1.1.3 Practical limitations of the attentional drift diffusion model

The aDDM incorporates the influence of gaze allocation on choice behavior through gaze-dependent drift rate changes in the decision process (see eq. 1.1-1.2). Due to these drift-rate changes, no analytical solution for the model's first passage time (FPT) distribution (which describes the likelihood that each item is chosen over time) exists. For this reason, fitting the aDDM to empirical response and gaze data requires a numerical estimate of its FPT, which can be obtained through model simulations. However, simulating the aDDM also requires a generative model of the underlying fixation process (for example, to continue the simulation after the empirically observed response time, where no empirical gaze data exists). Obtaining a fixation model might be simple for the two alternative case (as done in Krajbich et al., 2010), where the individual starts the decision process by looking with a certain probability at one alternative and then simply goes back and forth between the two until a decision is reached. However, the complexity of this problem drastically grows with an increasing choice set size (due to an exponential growth in the number of possible search trajectories) and a focus on the individual level (which requires individually tailored models of the fixation process). While these problems can be solved in theory, they are generally not of main interest to researchers studying simple choice behavior.

Due to these limitations, the application of the aDDM has been generally limited to small choice sets and group-level data, by estimating a single parameter set for an entire group of individuals (see, for example, Fisher, 2017, Krajbich et al., 2010, 2012, Krajbich and Rangel, 2011).

1.1.4 The need to study the individual

While group-level statistics can be informative for some research questions (for example, to specifically address differences between groups or experimental conditions or to forecast product sales in economic research), they can be unsuitable for understanding the choice behavior of an individual. Aggregate models can lead to false conclusions about true underlying individual processes (Grandy et al., 2017, Lewandowsky and Farrell, 2010). In a learning task, for example, the group-level average learning curve would appear as a gradual, smooth function over time, even if all individuals showed abrupt, step-like learning curves (much like an epiphany), but with variable learning onsets across individuals (Hayes, 1953). In this case, the group-level model would not accurately describe any individual of the group, and the deduction that individual learning occurs smoothly would be false. Similarly, using a single model parameter set to describe the choice behavior of a group could lead to false conclusions about the behavior of the underlying individuals. Therefore, it is crucial to study choice behavior at the level of the individual.

Previous applications of the aDDM specified a constant gaze bias for all individuals without rigorously testing the performance of the model at the level of the individual (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011). A rigorous test of gaze bias effects at the level of the individual should ideally be based on non-restricted individual model fits, include a comparison to models without gaze bias, and establish that the model provides an accurate account of the individually observed data. If, for example,

individual's decisions were affected differently by gaze behavior, we would find that the choices of some individuals would be more biased by gaze than the choices of others and thereby be more inconsistent with the values of the items in a choice set. Imagine, for example, a choice between two bottles of juice at the supermarket: one has a slightly higher value for the decision-maker than the other, but it is also less visually salient (e.g., Towal et al., 2013). If a person's association of gaze and choice behavior is strong, their choice would be biased towards the more visually salient bottle that attracts more of their gaze, even though it has a lower value. Conversely, if the person's association is weak, they would be able to select the higher valued option, despite their gaze being attracted more towards the visually salient but lower valued option. Accordingly, if the strength of this association is variable across individuals, it is necessary to account for these differences to accurately explain and predict individual choice behavior.

1.1.5 Many-alternative forced choice

While much effort has been devoted to understanding the mechanisms underlying two-alternative forced choice (2AFC) in value-based decision-making (Alós-Ferrer, 2018, Bhatia, 2013, Clithero, 2018, Hare et al., 2009, Hunt et al., 2018, Hutcherson et al., 2015, Krajbich et al., 2010, Martino et al., 2006, Milosavljevic et al., 2010, Philiastides and Ratcliff, 2013, Polanía et al., 2019, Rodriguez et al., 2014, Webb, 2018) and choices involving three to four alternatives (Berkowitsch et al., 2014, Boorman et al., 2013, Diederich, 2003, Gluth et al., 2020, 2018, Krajbich and Rangel, 2011, Noguchi and Stewart, 2014, Roe et al., 2001, Towal et al., 2013, Trueblood et al., 2014, Usher and McClelland, 2004), comparably little has been done to investigate many-alternative forced choices (MAFC, more than four alternatives; Ashby et al., 2016, Payne, 1976, Reutskaja et al., 2011).

Prior work on 2AFC has indicated that simple value-based choices are made through a process of gaze-driven evidence accumulation and comparison, as captured by the aDDM (for further details on the aDDM, see section 1.1.2). While this framework can in theory be extended to MAFC (see, for example, Krajbich and Rangel, 2011, Towal et al., 2013), it is still unknown whether it can account for choices from truly large choice sets.

In contrast, past research in MAFC suggests that people may resort to a "satisficing" strategy. Here, the idea is that people set a minimum threshold on what they are willing to accept and search through the alternatives until they find one that is above that threshold (McCall, 1970, Schwartz et al., 2002, Simon, 1955, 1956, 1957, 1959, Stüttgen et al., 2012). Satisficing has been observed in a variety of choice scenarios, including tasks with a large number of alternatives (Caplin et al., 2011, Stüttgen et al., 2012), patients with damage to the prefrontal cortex (Fellows, 2006), inferential decisions (Gigerenzer and Goldstein, 1996), survey questions (Krosnick, 1991), risky financial decisions (Fellner et al., 2009), and with increasing task complexity (Payne, 1976).

Past work has also investigated MAFC under strict time limits (Reutskaja et al., 2011). There, the authors compare the satisficing model to an optimal choice model where the decision-maker inspects as many items as possible in the available time and

then picks the best one (Rapoport and Tversky, 1966, Simon, 1955, 1959). They find that the best fitting model is in fact a hybrid of the two. What is unclear, however, is how well such a hybrid model performs compared to a model of gaze-driven evidence accumulation.

There is some empirical evidence that points towards a gaze-driven evidence accumulation and comparison process for MAFC. For instance, individuals look back and forth between alternatives as if comparing them (Russo and Rosen, 1975). Also, frequently looking at an item dramatically increases the probability of choosing that item (Chandon et al., 2009). First empirical evidence has also indicated that individuals use a gaze-dependent evidence accumulation process when making choices from sets of up to eight alternatives Ashby et al. (2016).

1.2 Research questions and hypotheses

Previous work has indicated that gaze allocation plays an active role in simple choices, such that items that are looked at longer generally have a higher probability to be chosen. This finding has been established on the group-level and in simple choice tasks with two to four alternatives (e.g., Armel et al., 2008, Cavanagh et al., 2014, Fiedler and Glöckner, 2012, Folke et al., 2016, Glöckner and Herbold, 2011, Konovalov and Krajbich, 2016, Krajbich et al., 2010, 2012, Krajbich and Rangel, 2011, Pärnamets et al., 2015, Shimojo et al., 2003, Stewart et al., 2016a,b, Vaidya and Fellows, 2015). It is not clear, however, whether I) this gaze bias effect is similarly present on the level of the individual, II) whether its strength is variable between individuals, and III) whether it generalizes to choice situations involving many alternatives.

The goal of this work is to explore these research questions, by testing the following set of hypotheses:

1. **Individuals exhibit a gaze bias in simple choice situations with few alternatives.** Previous work has demonstrated that gaze allocation and choice behavior are associated on the group-level and that models which account for this association perform well in capturing simple choice behavior (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011). We therefore hypothesize that this association is also present on the level of the individual in simple choice situations.

2. **The strength of the gaze bias varies between individuals.** First preliminary evidence has indicated that the strength of the association between gaze allocation and choice behavior is variable (c.f., Supplementary Fig. 11 in Krajbich et al., 2010). Further, a wealth of empirical findings indicates that many behavioural measures are generally variable between individuals (e.g., Grandy et al., 2017, Lewandowsky and Farrell, 2010, Ratcliff et al., 2006, 2010, Schmiedek et al., 2009). We therefore hypothesize that the strength of the association between gaze allocation and choice is also variable between individuals.

3. **Individuals with a strong gaze bias are generally less likely to choose the best item from a choice set compared to individuals with a weak gaze bias.** Previous work has established that longer gaze towards an item generally increases the probability that the item is chosen (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011). We would therefore expect that any variability in the strength of the association of gaze allocation and choice is directly linked to an individual's probability of choosing the best item from a choice set, such that stronger associations of gaze and choice correlate with lower probabilities of choosing the best item. The reasoning behind this is that a stronger association of gaze and choice could bias the decision process towards options that were looked at longer, even when they have a lower value.

4. **Gaze allocation and choice behavior exhibit a similar positive association in many-alternative choices as they do in smaller choice sets.** While the process of gaze-driven evidence accumulation and comparison, as captured

by the aDDM (section 1.1.2), can in theory be extended to MAFC, it is unclear whether it can actually account for individuals' behavior in these situations. First empirical evidence points towards a similar association of gaze allocation and choice in MAFC, as frequently looking at an item dramatically increases the probability of choosing that item (Chandon et al., 2009), and individuals seem to use a gaze-dependent evidence accumulation process when making choices from sets of up to eight alternatives (Ashby et al., 2016). For this reason, we hypothesize that gaze allocation and choice behavior are similarly linked in MAFC as they are in choices from smaller choice sets.

1.3 Methodology

1.3.1 Overview of datasets and experiment tasks

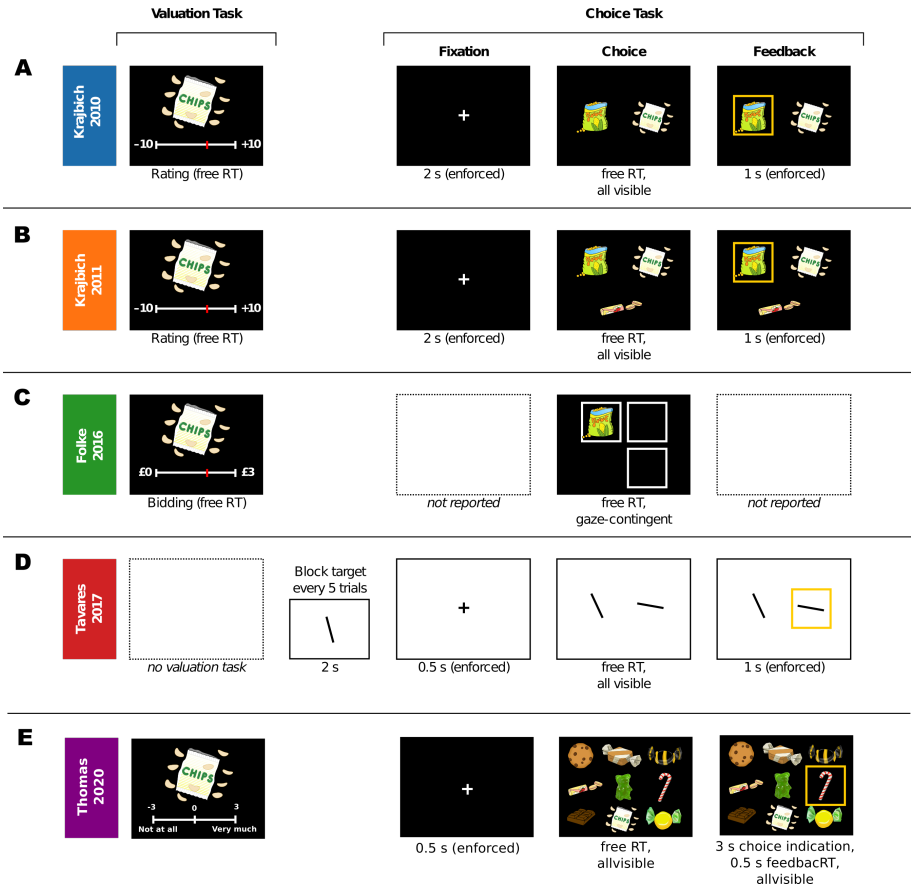

Figure 1.1: Experimental paradigms. A-E: We included five datasets in our analyses. These include four value-based (Krajbich 2010 (A), Krajbich 2011 (B), Folke 2016 (C), Thomas 2020 (E)) and one perceptual choice experiment (Tavares 2017 (D)). In all experiments, participants were instructed to choose the best out of two (A, D), three (B, C), or 9, 16, 25, or 36 items (E) (that is, the item they would like to eat most in value-based tasks or the item most similar to a target stimulus that was presented every five trials in the perceptual task). Three of the value-based experiments included a valuation task before the main choice task, whereby participants either rated each item (A, B) or indicated their willingness-to-pay in a Becker–DeGroot–Marschak procedure (C). The valuation task of the fourth value-based experiment (Thomas 2020; E) was performed after the main choice task. All choices were made without time restrictions. The choice task in Folke 2016 (C) used a gaze-contingent presentation, whereby items were only revealed when the participant's gaze was directed to an item's location on the screen. Experiments used real snack food items instead of illustrations. For additional details, see Appendix A.1.1.

We investigated the relation between gaze allocation and choice behavior across five datasets (for further details, see Fig. 1.1, Appendix A.1.1, and Thomas et al., 2020,

2019b). In each dataset, healthy participants made repeated decisions between multiple options while their eye movements, choices, and response times were recorded.

Note that we define a gaze to an item as all consecutive fixations towards the item that happen without any interrupting fixation to other parts of the choice screen (for further details on the preprocessing of eye movement data, see Appendix A.1.1).

The first dataset was published by Krajbich and colleagues in 2010 (Krajbich et al., 2010) (henceforth referred to as Krajbich 2010). In the corresponding experiment, hungry participants were asked to choose the item that they would like to eat most from a set of two snack food items without any time restrictions (see Fig. 1.1 A). Participants also gave a liking rating for each of the 70 snack food items that were used in the experiment, indicating how much they would like to eat the item at the end of the experiment. This dataset includes 39 participants, each of whom performed 100 trials.

The second dataset was published by Krajbich and Rangel in 2011 (Krajbich and Rangel, 2011) (henceforth referred to as Krajbich 2011) and is similar to Krajbich 2010. In Krajbich 2011, participants chose between three snack food items that were arranged in a triangular shape (see Fig. 1.1 B). As in Krajbich 2010, participants provided liking ratings for all available items in a separate task. This dataset includes 30 participants, each of whom performed 100 trials.

The third dataset consists of experiment 2 from a study that was published by Folke and colleagues in 2016 (Folke et al., 2016) (henceforth referred to as Folke 2016). In this experiment, 24 hungry participants performed 144 trials of a task that closely resembled the Krajbich 2011 three-alternative forced-choice snack food task (see Fig. 1.1 C). Unlike in Krajbich 2011, however, the choice task used a gaze-contingent presentation, whereby items were only revealed when the participant's gaze was directed to an item's location on the screen. In addition, after each choice, the participants provided confidence ratings (which we did not use in this study). Similar to Krajbich 2010 and Krajbich 2011, item values were estimated in a separate task, in which a Becker-DeGroot-Marschak auction procedure was used to elicit willingness-to-pay estimates (Becker et al., 1964).

The fourth dataset is experiment 1 from a study by Tavares and colleagues that was published in 2017 (Tavares et al., 2017) (henceforth referred to as Tavares 2017). This dataset is qualitatively different from the other three value-based choice datasets. Participants made perceptual judgments about the orientations of two line segments and were asked to decide which is closer to a target (see Fig. 1.1 D). In this case, we define the value of an item by its angular distance to the target (with higher values for smaller distances). This dataset includes 25 participants, each of whom performed 1,344 trials across four sessions.

The fifth dataset was collected by my colleagues and me (Thomas et al., 2020) (henceforth referred to as Thomas 2020). This dataset is very similar to Krajbich 2010 and Krajbich 2011, but extends the choice set size to many alternatives. Specifically, in this experiment hungry participants were asked to choose the item that they would like to eat most from sets of 9, 16, 25, and 36 snack food items. As in Krajbich 2010 and Krajbich 2011, participants provided liking ratings for all available items in a separate liking rating task (see Fig. 1.1 E). This dataset includes 49 participants, each of whom

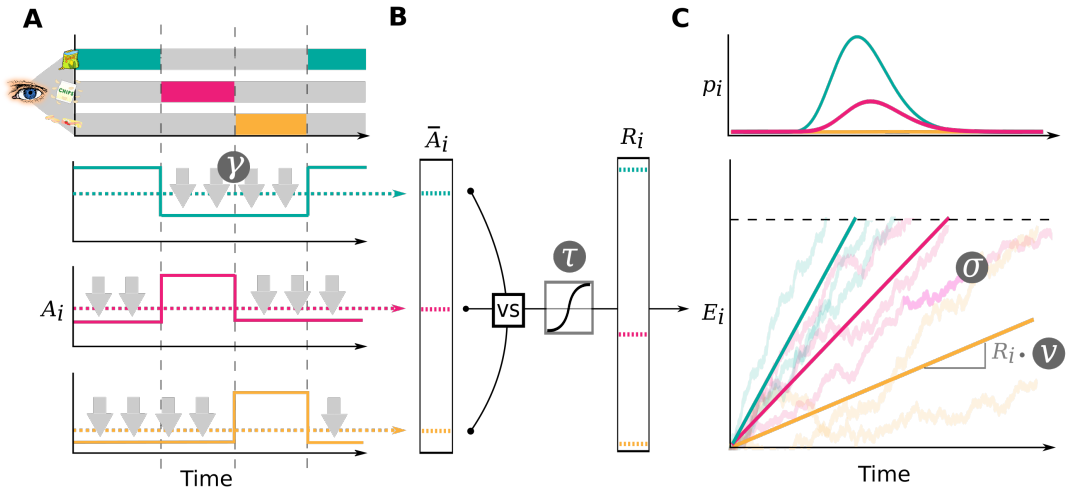

Figure 1.2: Gaze-weighted linear accumulator model (GLAM). In the GLAM, preference formation during the decision process is dependent on the allocation of visual gaze (A) and driven by two decision signals: an absolute and relative decision signal. The magnitude of the absolute decision signal A_i is determined by the momentary allocation of visual gaze: While an item is currently not looked at, its signal is discounted by γ (discounting is illustrated by gray arrows in A). Relative decision signals R_i for each item in the choice set are computed in three steps: First, an average absolute decision signal $\overline{A}_i$ for each item i is computed (dashed lines in A). Second, the difference between each average absolute decision signal and the maximum of all others is determined (B). Third, the resulting differences are scaled through a logistic transform, as the GLAM assumes an adaptive representation of the relative decision signals that is especially sensitive to differences close to 0 (where the absolute signal for an item is very close to the maximum of all others). The resulting relative decision signals R_i can be used to predict choice and response time, by determining the speed of the accumulation process in a linear stochastic race (C). The stochastic race provides first-passage time distributions p_i, describing the likelihood of each item being chosen at each time point.

performed 50 trials for each of the four choice set sizes.

Overall, the five datasets span 167 individuals, six choice set sizes (2, 3, 9, 16, 25, and 36 alternatives), and two choice domains (value-based and perceptual choice).

1.3.2 The gaze-weighted linear accumulator model (GLAM)

To investigate the association of gaze allocation and choice behavior on the level of the individual and in choice situations with many alternatives, we propose the gaze-weighted linear accumulator model (GLAM). The GLAM is inspired by the aDDM (see section 1.1.2 and Krajbich and Rangel, 2011) and assumes that preference formation, during a simple choice process, is guided by the allocation of visual gaze. Particularly, the decision process is guided by a set of two decision signals (see Fig. 1.2): An absolute and relative decision signal. While a decision is being formed, the absolute decision signal A_i of an item i can be in two states (Fig. 1.2 A): An unbiased state, equal to the item's value r_i while the item is looked at, and a biased state while any other item

is looked at, where the item value r_i is discounted by γ. The average absolute decision signal $\overline{A}_i$ (Fig. 1.2 B) is then given by:

$$\overline{A}_i = g_i \cdot r_i + (1 - g_i) \cdot \gamma \cdot r_i, \tag{1.5}$$

where g_i is defined as the fraction of total trial time that item i was looked at. If $\gamma = 1$, there is no difference between the biased and unbiased state, resulting in no influence of gaze allocation on choice behavior. For $0 < \gamma < 1$, the absolute decision signal A_i is discounted, resulting in generally higher choice probabilities for items that have been looked at longer. For $\gamma < 0$, the sign of the absolute decision signal A_i changes when the item is not looked at, leading to an overall even stronger gaze bias as evidence for these items is actively lost when they are not looked at. This type of gaze-dependent leakage mechanism is supported by a variety of recent empirical findings (Ashby et al., 2016, Krajbich et al., 2012, Lopez-Persem et al., 2016, Smith and Krajbich, 2018, Thomas et al., 2019b, Trueblood et al., 2013).

Importantly, recent empirical work has also indicated another formulation of the gaze bias mechanism (Cavanagh et al., 2014, Westbrook et al., 2020), in which gaze plays an additive role in the decision process and thereby increases the probability that an item is chosen, independently of the item's value r_i. To account for this alternative hypothesis of the functional form of the gaze bias mechanism, two other definitions of the average absolute decision signal A_i will be used over the course of this work:

$$\overline{A}_i = \begin{cases} g_i \cdot (r_i + \zeta) + (1 - g_i) \cdot r_i, & : \text{additive gaze bias variant} \\ g_i \cdot (r_i + \zeta) + (1 - g_i) \cdot \gamma \cdot r_i, & : \text{full gaze bias variant} \end{cases} \tag{1.6}$$

These variants differ from the regular GLAM formulation in that they add a constant ζ ($0 \leqslant \zeta$) to the value r_i of an item, when the item is looked at, thereby, biasing the decision process towards the item, independent of the item's value.

To determine the relative decision signals (Fig. 1.2 B), the average absolute decision signals $\overline{A}_i$ are transformed in two steps: First, for each item i, the relative evidence R_i^* is computed as the difference between the average absolute decision signal of the item $\overline{A}_i$ (eq.1.5) and the maximum of all other average absolute decision signals $\overline{A}_j$ (also obtained from eq.1.5):

$$R_i^* = \overline{A}_i - max_{j \neq i} \overline{A}_j \tag{1.7}$$

Second, the relative evidences R_i^* are scaled through a logistic transform to obtain the relative decision signals R_i:

$$R_i = \sigma(R_i^*) \tag{1.8}$$

$$\sigma(x) = \frac{1}{1 + e^{(-\tau \cdot x)}} \tag{1.9}$$

The GLAM assumes an adaptive representation of the relative decision signals R_i, which is maximally sensitive to small differences in the relative evidences R_i^* close to 0 (where the difference between the average absolute decision signal $\overline{A}_i$ of an item and the maximum of all others is small).

The sensitivity of this transform is determined by the temperature parameter τ of the logistic function. Larger values of τ indicate stronger sensitivity to small differences in the average absolute decision signals $\overline{A}_i$.

To capture an individual's choice behavior and response time (RT), the GLAM employs a linear stochastic race, with one evidence accumulator E_i for each choice alternative i (Fig. 1.2 C):

$$E_i(t) = E_i(t-1) + v \cdot R_i + \mathcal{N}(0, \sigma^2) \text{ with } E(0) = 0 \tag{1.10}$$

At each time step t, the amount of accumulated evidence is determined by the accumulation rate $v \cdot R_i$, and zero-centered normally distributed noise with standard deviation σ. The velocity parameter v linearly scales the relative decision signals in the race process and thereby affects the response times produced by the model: Lower values of v produce longer response times, while larger values of v result in shorter response times. A choice for an item is made as soon as one accumulator reaches the decision boundary b. To avoid underdetermination of the model, either the velocity parameter v, the noise parameter σ, or the boundary b has to be fixed. Similar to the aDDM, the GLAM fixes the boundary to a value of 1. The first passage time density $f_i(t)$ of a single linear stochastic accumulator E_i, with decision boundary b, is given by the inverse Gaussian distribution:

$$f_i(t) = \Big[\frac{\lambda}{2\pi t^3}\Big]^{\frac{1}{2}} \exp\Big(\frac{-\lambda(t-\mu)^2}{2\mu^2 t}\Big) \tag{1.11}$$

$$\text{with } \mu = \frac{b}{vR_i} \text{ and } \lambda = \frac{b^2}{\sigma^2}$$

However, this density does not take into account that there are multiple accumulators in each trial racing towards the same boundary. For this reason, $f_i(t)$ must be corrected for the probability that any other accumulator crosses the boundary first. The probability that an accumulator crosses the boundary prior to t, is given by its cumulative distribution function $F_i(t)$:

$$F_i(t) = \Phi\Big(\sqrt{\frac{\lambda}{t}}\big(\frac{t}{\mu} - 1\big)\Big) + \exp\Big(\frac{2\lambda}{\mu}\Big) \cdot \Phi\Big(-\sqrt{\frac{\lambda}{t}}\big(\frac{t}{\mu} + 1\big)\Big) \tag{1.12}$$

Here, $\Phi(x)$ defines the standard normal cumulative distribution function. Hence, the joint probability $p_i(t)$ that accumulator E_i crosses b at time t, and that no other accumulator $E_{j\neq i}$ has reached b first, is given by:

$$p_i(t) = f_i(t) \prod_{j \neq i} \big(1 - F_j(t)\big) \tag{1.13}$$

Importantly, all of the GLAM's parameters can be recovered to a satisfying degree without bias (see Appendix Fig. A.2).

1.3.3 GLAMbox

To make the GLAM easily accessible for other researchers, we have built a Python-based toolbox, called GLAMbox (Molter et al., 2019), which is built upon PyMC3 (Salvatier et al., 2016). GLAMbox enables Bayesian parameter estimation of the GLAM for individual, pooled or hierarchical models, provides an easy-to-use interface to predict choice behaviour and visualize choice data, and benefits from all of PyMC3's Bayesian statistical modeling functionality. Further documentation, resources and the toolbox itself are available at https://glambox.readthedocs.io.

1.3.4 Competing models for many-alternative choices

The following describes the set of models that we will compare in the many-alternative choice dataset (Thomas 2020; for further details on the dataset, see section 1.3.1 and Thomas et al., 2020). These models span the space between traditional models of rational choice and gaze-driven evidence accumulation. The optimal choice model with zero search costs as well as the hard and probabilistic satisficing models are based on a proposal by Reutskaja and colleagues (Reutskaja et al., 2011). The independent evidence accumulation is closely related to the GLAM, but assumes that items are evaluated independently from the others.

Optimal choice with zero search costs: The optimal choice model with zero search costs is based on the framework of rational decision-making (Luce and Raiffa, 1957, Simon, 1955) and assumes that individuals first look at all the items in a choice set and then choose according to the following choice rule: with probability β, the model chooses the item with the highest seen value in the trial. On the other hand, with probability $1-\beta$, the model makes a probabilistic choice over the set of seen items according to a softmax function with scaling parameter τ. Here, item i (with value r_i) is chosen with probability σ_i:

$$\sigma_i = \frac{e^{\tau \cdot r_i}}{\sum_i e^{\tau \cdot r_i}} \tag{1.14}$$

Hard satisficing: The hard satisficing model assumes that individuals search until either all the items of a choice set have been seen, or they find an item with reservation value V or higher (Caplin et al., 2011, Fellows, 2006, McCall, 1970, Payne, 1976, Schwartz et al., 2002, Simon, 1955, 1956, 1957, 1959, Stüttgen et al., 2012). In the former case, individuals make a probabilistic choice over the set of seen items, as in the optimal choice model. In the latter case, individuals immediately stop their search and choose the first item that meets the reservation value. Crucially, the reservation value can vary across individuals and choice set size conditions.

Probabilistic satisficing: The probabilistic satisficing model is based on the hybrid model that was proposed by Reutskaja and colleagues (Reutskaja et al., 2011), which combines elements from optimal choice and hard satisficing, but extends this model by the addition of a gaze bias mechanism. Specifically, the probabilistic satisficing model

is guided by two rules: a probabilistic stopping rule, defining the probability $q(t)$ with which the search ends and a choice is made at each time point t (with $\Delta t = 1ms$), and a probabilistic choice rule, defining a choice probability p_i for each item i in the choice set. At each point in time, the stopping probability $q(t)$ increases with the amount of elapsed time in the trial and the cached (i.e., highest-seen) value $C(t)$.

$$q(t) = v \cdot t + \alpha \cdot C(t), \text{ with } 0 \leqslant \text{q(t)} \leqslant 1 \tag{1.15}$$

The strength of the influence of $C(t)$ and t on $q(t)$ is determined by the two positive linear weighting parameters α and v. Importantly, the cached value $C(t)$ is dependent on gaze allocation:

$$C(t) = \max_I c_i(t) \tag{1.16}$$

$$c_i(t) = g_i(t) \cdot (r_i + \zeta) + (1 - g_i(t)) \cdot \gamma \cdot r_i \tag{1.17}$$

Here, $g(t)$ is defined as the fraction of cumulative time t that an item was looked at, while γ and ζ determine the strength of the gaze bias effect: when an item is looked at, its value r_i is increased by a positive constant ζ ($0 \leqslant \zeta$), while the item's value is discounted by γ ($0 \leqslant \gamma \leqslant 1$), when the item is momentarily not looked at. In the probabilistic satisficing model, gaze allocation thereby increases the value of an item, which influences the stopping and choice probabilities. Note that by setting $\gamma = 1$ and $\zeta = 0$ the gaze bias mechanisms can be entirely removed from the model.

By itself, $q(t)$ does not account for the probability that the search has ended at any time point prior to t. In order to apply and fit the model to response time data, we need to compute the joint probability $f(t)$ that the search has not stopped prior to t and the probability that the search ends at time point t. Therefore, we correct $q(t)$ for the probability $Q(t)$ that the search has not stopped at any time point prior to t:

$$f(t) = q(t) \cdot Q(t-1) \tag{1.18}$$

$$Q(t) = \prod_1^t (1 - q(t)) \tag{1.19}$$

Once the search process ends, the model makes a probabilistic choice over the set of seen items, following a softmax choice rule (with scaling parameter τ) of their gaze-weighted values $c(t)$:

$$\sigma_i(t) = \frac{e^{\tau \cdot c_i(t)}}{\sum_I e^{\tau \cdot c_i(t)}} \tag{1.20}$$

Lastly, by multiplying the joint probability $f(t)$ by $\sigma_i(t)$, we obtain the probability $p_i(t)$ that item i is chosen at time point t:

$$p_i(t) = f(t) \cdot \sigma_i(t) \tag{1.21}$$

Independent evidence accumulation: Like the GLAM, the independent evidence accumulation model belongs to the class of linear stochastic race models (see section 1.3.2 and Thomas et al., 2020). In contrast to the GLAM, the independent evidence accumulation model assumes that evidence accumulation for an item is independent of all the other items in a choice set and only starts once the item was looked at in a trial.

Specifically, the decision process is guided by an absolute decision signals A_i for each item i that was looked at in a trial. The absolute decision signal A_i implements the model's gaze bias mechanisms and can be in two states: An additive state, in which the item's value r_i is amplified by the addition of ζ ($0 \leqslant \zeta$), while the item is looked at, and a multiplicative state while any other item is looked at, in which the item's value r_i is discounted by γ ($0 \leqslant \gamma \leqslant 1$).

$$A_i(t) = \begin{cases} r_i + \zeta, & \text{if } i \text{ is looked at} \\ r_i \cdot \gamma, & \text{otherwise} \end{cases} \tag{1.22}$$

Note that by setting $\gamma = 1$ and $\zeta = 0$ the model's gaze bias mechanisms can be entirely removed from the model. Similar to the GLAM (see eq. 1.5), the average absolute decision signal $\overline{A}_i$ is obtained by:

$$\overline{A}_i = g_i \cdot (r_i + \zeta) + (1 - g_i) \cdot \gamma \cdot r_i \tag{1.23}$$

Here, g_i indicates the fraction of remaining trial time, after item i was first looked at in a trial, that the individual spent looking at item i. If an item was not looked at in a trial, we set $\overline{A}_i = 0$.

At each time step t, the amount of accumulated evidence is determined by a velocity parameter v, the item's average absolute decision signal $\overline{A}_i$, and zero-centered normally distributed noise with standard deviation σ:

$$E_i(t) = E_i(t-1) + v \cdot \overline{A}_i + \mathcal{N}(0, \sigma^2) \tag{1.24}$$

Once any of the evidence accumulators E_i reaches a common decision boundary b (which we set to 1), a choice for the respective item is made. Note that we set $E_i(t < t0_i) = 0$ for all items that were not yet looked at by time point t ($t0_i$ indicates the time point in the trial at which item i was first seen).

As for the GLAM, the joint probability $p_i(t)$ that accumulator E_i crosses boundary b at time t, and that no other accumulator $E_{j \neq i}$ has reached b first, can be obtained according to eq. 1.13, when accounting for the different accumulation onsets $t0_i$.

1.3.5 Mixed-effects modeling

All regression coefficients reported throughout represent fixed effects from Bayesian mixed-effects linear (for continuous dependent variables) or logistic (for binary dependent variables) regression models, including a random intercept and slope for each dataset on each predictor. For each fixed effect, the posterior mean (β) and the associated 95% highest posterior density interval (HDI) values are reported (see Appendix A.1.4 for further details).

1.4 Empirical results

1.4.1 Study I: Individual gaze bias differences

First, we investigated whether the previously reported group-level link between gaze allocation and choice behavior in small choice sets (Krajbich et al., 2010, Krajbich and Rangel, 2011) exists on the level of the individual and whether its strength is variable between individuals.

To this end, we analyzed the choice, response time, and gaze data of four previously published simple choice experiments (namely, Krajbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017; for an overview of the datasets, see Section 1.3.1 and Appendix A.1.1). In total, the four datasets include 118 individuals, two choice set sizes (two- and three-alternative), and two choice domains (value-based and perceptual).

We first studied general behavioral differences between the individuals in the data on the following three behavioral metrics: participants' mean response time (RT), mean probability of choosing the best item (we defined the best item either as the item with the highest liking rating or willingness-to-pay in the value-based choice tasks, or the item with the smaller angular distance to the target in the perceptual choice task, see Fig. 1.1, section 1.3.1 and Appendix A.1), and influence of gaze allocation on choice probability (defined as the mean increase in choice probability for an item that was looked at longer than the others, after correcting for the influence of the item's value or angular distance on choice probability; for further details on this measure, see Appendix A.1).

Participants differed considerably in all three behavioral metrics (Fig. 1.3). Notably, 98% of the participants also showed positive scores on the gaze influence measure, indicating an overall positive relationship between gaze allocation and choice probability. Importantly, individual scores on this measure varied substantially and ranged from -11% to 72% (Fig. 1.3 B).

We also probed the relationship between the behavioral metrics, to better understand their dynamics. We did not find any association between participants' probability of choosing the best item and their mean RT ($\beta = -0.19\%$, 95% HDI = [-3.08%, 2.87%] per second increase in RT; Fig. 1.3) D) and no association between participants' mean RT and gaze influence ($\beta = -1$ ms, 95% HDI = [-33 ms, 33 ms]) per percentage increase in the gaze influence measure; Fig. 1.3 E). However, participants' probability of choosing the best item from a choice set decreased with increasing individual gaze influence measures ($\beta = -0.34\%$, 95% HDI = [-0.71%, 0.08%] per percentage increase in the gaze influence measure, 95.2% of posterior density below 0; Fig. 1.3 F).

The four datasets strongly differed on the three behavioral metrics (see Appendix A.1). Yet, these differences between datasets cannot be attributed to the effect of choice domain (perceptual versus value-based) or set size (two versus three items) alone, as the original tasks also differed in other aspects (for example, different stimuli in value-based versus perceptual tasks, different number of trials and different presentation format). For this reason, we refrain here from interpreting these differences between datasets further.

The behavioral and eye-tracking data suggests substantial variability in the extent to which gaze affects individuals' choice behavior (Fig. 1.3 B). To provide conclusive

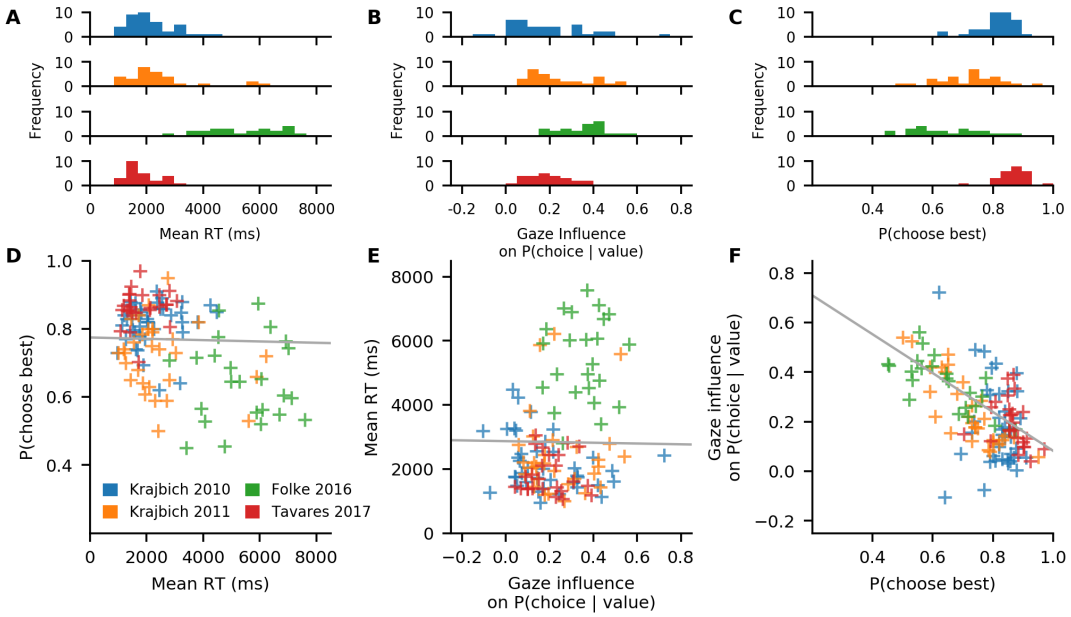

Figure 1.3: Individual differences in the three studied behavioral metrics and their associations in the choice sets with two to three alternatives. A–C: Distributions of individuals' mean RT (A), gaze influence (mean increase in choice probability for an item that is looked at longer than the others, after correcting for the influence of item value) (B), and probability of choosing the best item (C) per dataset. D: There is no association between the mean RT and the individual probability of choosing the best item ($\beta = -0.19\%$, 95% HDI = [-3.08%, 2.87%] per second increase in RT). E: There is no association between gaze influence and the mean RT ($\beta = -1$ ms, 95% HDI = [-33 ms, 33 ms] per percentage increase in the gaze influence measure). F: An individual's probability of choosing the best item decreases with increasing gaze influence ($\beta = -0.34\%$, 95% HDI = [-0.71%, 0.08%] per percentage increase in the gaze influence measure, 95.2% of posterior density below 0). Each cross represents one individual participant. Grey lines represent the fixed effect from mixed-effects regression models with random slopes and intercepts for each dataset.

quantitative evidence for or against the presence of gaze bias mechanism on the level of the individual, we adopted a computational modeling approach, by fitting and comparing two GLAM variants (see section 1.3.2) to the RT, choice, and gaze data of each participant. The first is a GLAM variant with gaze bias and free parameters v, γ, σ and τ. This model allowed the gaze bias parameter γ to vary freely between individuals. The second is a no-gaze-bias GLAM variant, whereby the gaze bias parameter γ was fixed to 1 (resulting in no influence of gaze on the accumulation process). Overall, the GLAM variant with gaze bias fitted 109 out of 118 (92%) participants better than the no-gaze-bias variant[1]. Within each dataset, the data of 79% (Krajbich 2010), 97% (Krajbich 2011), 100% (Folke 2016) and 100% (Tavares 2017) of the participants were better described by the GLAM variant with gaze bias (Fig. 1.4 B).

[1]The two model variants differ in their complexity. The GLAM has one more free parameter and can therefore be expected to provide a better absolute fit to the data. To account for this, we used the

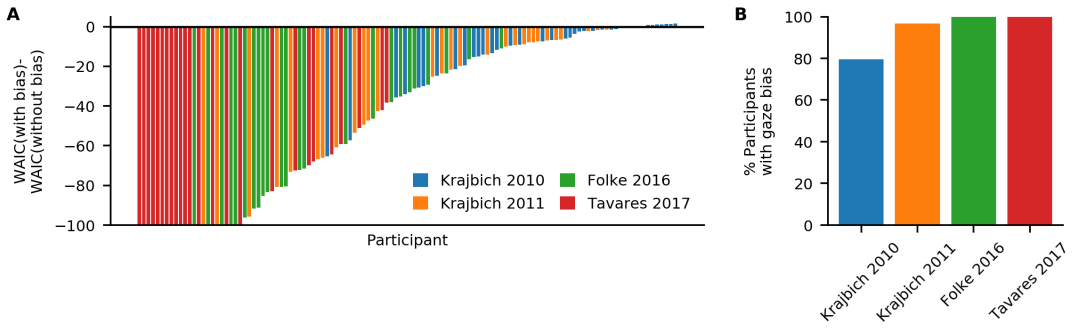

Figure 1.4: Individual relative model comparison between the GLAM and a restricted GLAM variant with no gaze bias in the choice sets with two to three alternatives. A: Individual WAIC differences between the GLAM and the GLAM variant without gaze bias. Negative differences indicate better fits of the GLAM. B: Across datasets, the response behavior of most individuals is better described by the GLAM (given by the lowest score on the WAIC). Note that the y-axis in A is truncated to better show small differences. The lowest WAIC difference was -400.64.

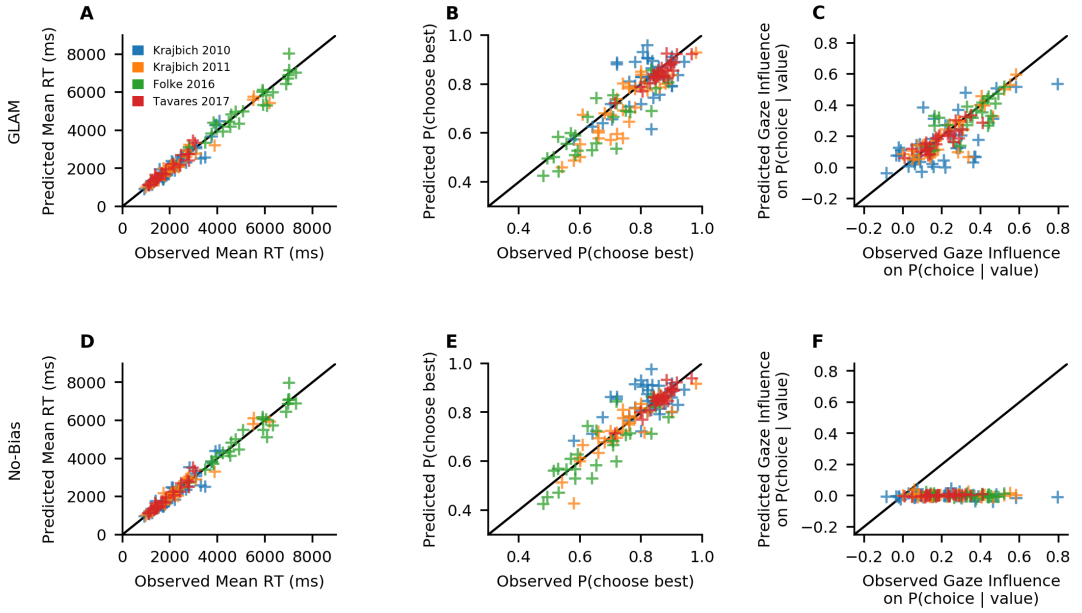

Figure 1.5: Individual out-of-sample predictions of behavioral metrics for all odd-numbered trials in the choice sets with two to three alternatives. A–C: The GLAM accurately predicts individuals' mean RT (A), probability of choosing the best item (B) and influence of gaze on choice probability (C). D–F: While the no-gaze-bias variant also accurately predicts individuals' mean RT (D) and probability of choosing the best item (E), it fails to accurately capture individuals' influence of gaze on choice probability (F). Each cross represents one individual participant. Model predictions are simulated using parameter estimates obtained from individual fits on even-numbered trials (for more details on the simulation procedures, see Appendix A.1.8)

widely applicable information criterion (WAIC) (Vehtari et al., 2017) to perform model comparisons at the level of the individual, as it includes a penalty for model complexity. Lower WAIC scores indicate a better model fit.

However, this analysis did not take into account whether the GLAM also accurately predicts individuals' behaviour on an absolute level. To test this, we used both model variants to simulate response data for each individual. In this analysis, we split the data into even- and odd-numbered trials. We then used all even trials to estimate individual model parameters (see Appendix A.1.5 for an overview of the estimation procedures). Subsequently, we predicted choices and RTs for all odd-numbered trials, thereby comparing model predictions to data that did not inform the parameter estimates (see Appendix A.1.8 for an overview of the simulation procedures). We note, however, that even- and odd-numbered trials from the same participant are not fully independent from one another. To assess the quality of the fit of both models' predictions to the empirically observed data across datasets, we performed the following test: For each model and each behavioural measure, we computed a mixed-effects regression, regressing the respective measure onto a binary variable, which indicates whether each value on this measure comes from the empirically observed data or from the model simulations. If the fixed-effects estimate of the indicator variable differed from 0, model predictions deviate meaningfully from observed data across datasets. Overall, the GLAM accurately predicted participants' RTs (β = -9 ms, 95% HDI = [-410, 344] difference between the observed and predicted data; Fig. 1.5 A), probability of choosing the best item (β = -2.22%, 95% HDI = [-7.03, 2.30] difference between the observed and predicted data; Fig. 1.5 B) as well as the strength of their gaze influence (β = -2.20%, 95% HDI = [-8.25, 4.21] difference between the observed and predicted data; Fig. 1.5 C). The variant without gaze bias predicted participants' individual mean RTs (β = 15 ms, 95% HDI = [-355, 413] difference between the observed and predicted data; Fig. 1.5 D) and the probability of choosing the best item (β = 0.13%, 95% HDI = [-4.43, 4.83] difference between the observed and predicted data; Fig. 1.5 E) similarly well. However, the restricted model by design cannot predict the influence of gaze on the participants' choices, resulting in no association between the predicted and empirical data in our gaze influence measure (β = -22.72%, 95% HDI = [-30.81, -13.20] difference between the observed and predicted data; Fig. 1.5 F).

Note that the GLAM further accurately recovered the observed associations between the three behavioral metrics as well as the distribution of RTs within and across individuals (Appendix Fig. A.3 and A.4).

We also compared the gaze bias mechanism implemented in the GLAM against a variant with an additive effect of gaze on choice behavior (see eq. 1.6 and Appendix Fig. A.5). This comparison revealed that similar proportions of participants were better described by either model variant, with an additional group of participants whose choice behavior was described similarly well by both variants. We therefore concluded that there is not a winning gaze bias mechanism that we can identify across individuals. Model simulations also revealed that both variants described participants' response behavior similarly well and mimicked each other considerably in their predictions, both on the individual (Appendix Fig. A.5 B-D) and group-averaged level (Appendix Fig. A.6). Further, the gaze bias estimates of both variants were highly correlated (Spearman's $\rho(117) = -0.86, P < 0.001$).

Next, we tested whether the GLAM's parameters are able to explain variability in participants' choice behavior (Fig. 1.6). We found that: v (velocity parameter)

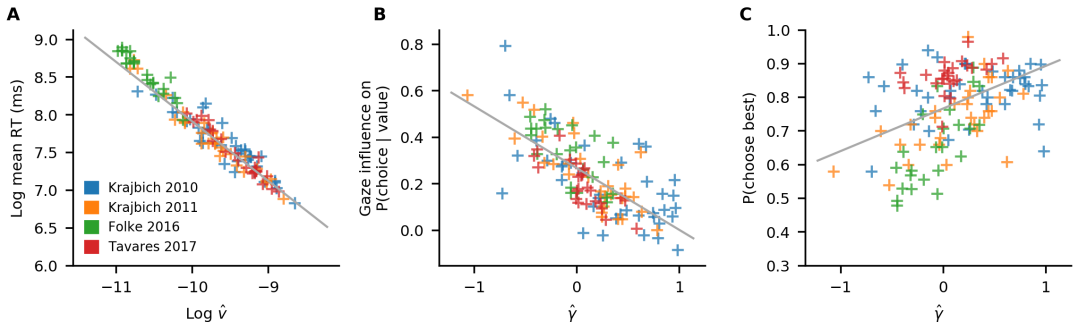

Figure 1.6: Associations between individuals' response behavior in the odd-numbered trials and the model parameters estimated from the even-numbered trials in the choice sets with two to three alternatives. A: log-transformed mean RTs decrease with increasing log-transformed v estimates ($\beta = -0.79$ log(ms), 95% HDI = [-0.85, -0.71] per unit increase in log(v)). B: Behavioral gaze influence decreases with increasing γ estimates ($\beta = -26.59\%$, 95% HDI = [-37.10, -17.24] per unit increase in γ). C: Individuals' probability of choosing the best item increases with increasing γ estimates ($\beta = 12.65\%$, 95% HDI = [-3.19, 28.84] per unit increase in γ; 94.5% of the posterior density estimates were greater than 0). Each cross represents one individual participant.

estimates scaled logarithmically with the participants' mean RT ($\beta = -0.79$ log(ms), 95% HDI = [-0.85, -0.71] per unit increase in log(v); Fig. 1.6 A), while γ (gaze bias parameter) estimates correlated with the strength of participants' gaze influence on choice probability ($\beta = -26.59\%$, 95% HDI = [-37.10, -17.24] per unit increase in γ; Fig. 1.6 B) and probability of choosing the best item ($\beta = 12.65\%$, 95% HDI = [-3.19, 28.84] per unit increase in γ; 94.5% of the posterior density estimates were greater than 0; Fig. 1.6 C).

Taken together, these findings indicate an active role of gaze allocation in the decision process of the individual: On the behavioral level, participants exhibited an overall positive relationship between gaze and choice (with longer gaze increasing choice probability). Similarly, in a structured model comparison, the response behavior of the majority of participants was better captured by a model with gaze bias mechanism than by one without. Importantly, the strength of the association between gaze allocation and choice (behaviorally and in model estimates, see Appendix A.1) was highly variable across individuals. Accounting for this variability was necessary in order to accurately predict individuals' response behaviour.

We also found that individuals' gaze bias estimates were predictive of their probability of choosing the best item from a choice set (stronger gaze biases were associated with more choices that were inconsistent with item values). This relationship can be explained as follows: the gaze bias parameter allows GLAM to bias the choice process according to the distribution of gaze between items. That is, with a strong gaze bias, GLAM's predictions are strongly dependent on the distribution of gaze, and a gaze distribution that is random with respect to the value of the items then leads to more random choices. Conversely, GLAM's predictions are independent of gaze when no gaze bias is present. GLAM then neglects the gaze data and predicts choices solely driven

by the values of the items. The strength of individuals' gaze bias thereby represents another source of variability in individuals' ability to choose the best item from a choice set. This variability was often attributed to differences in generic accumulation noise parameters (Ratcliff et al., 2006, 2010), obscuring further insight into the mechanisms driving these individual differences.

1.4.2 Study II: Choices from many alternatives

Choice behavior

Our findings have demonstrated that gaze allocation plays an active role in the decision process of the individual in choice situations with two to three alternatives. It is unclear, however, whether gaze allocation plays a similarly active role in the decision process when individuals make choices from many alternatives ($N \geqslant 4$). In contrast, research in many-alternative choice (MAFC) has often argued for models of optimal choice, satisficing, or hybrids of the two (McCall, 1970, Reutskaja et al., 2011, Schwartz et al., 2002, Simon, 1955, 1956, 1957, 1959, Stüttgen et al., 2012)).

To bridge this divide, we developed and compared a set of models between optimal choice and gaze-driven evidence accumulation (for an overview of the models, see section 1.3.4) on choice, RT, and gaze data of 49 individuals who performed a simple value-based choice task involving sets of 9, 16, 25, and 36 snack foods (Thomas 2020; for details on the task and data, see Section 1.3.1 and Appendix A.1.1).

First, we probed the assumptions of the optimal choice model with zero search costs (see Section 1.3.4), which assumes that individuals look at all the items in a choice set before making a choice and then choose the highest-rated item at a fixed rate. Conditional on the set of looked-at items, subjects chose the highest-rated item at a very consistent rate across set sizes, with an average of 84% (β = 0.05%, 95% HDI = [-0.04,0.14] per item; Fig. 1.7 A), in line with the assumption of the optimal choice model. Nevertheless, subjects did not look at all food items in a given trial (Fig. 1.7 B), while the fraction of items in a choice set that subjects looked at decreased with choice set size (β = -1.5%, 95% HDI = [-1.5, -1.4] per item; Fig. 1.7 B) and mean RTs increased (β = 85 ms, 95% HDI = [67, 102] per item; Fig. 1.7 C). This immediately ruled out a strict interpretation of the optimal choice model, as subjects did not look at all items before making a choice.

Next, we tested the assumption of satisficing choice. Specifically, we compared two variants of satisficing choice, with a hard and probabilistic stopping rule (for further methodological details, see section 1.3.4). The hard satisficing model predicts that subjects should stop their search and make a choice as soon as they find an item that meets their acceptance threshold. Accordingly, the last item that subjects look at should be the one that they choose (unless they look at every item). However, across choice set sizes, subjects only chose the last item that they looked at in 48.6% of the trials (β = 0.13%, 95% HDI = [-0.001, 0.26] per item; Fig. 1.7 D), contradicting the hard satisficing model. Even within the trials where subjects did not look at every item, the probability that they chose the last seen item was on average only 48.4%. The probabilistic satisficing model, on the other hand, predicts that the probability with which subjects stop their search and make a choice generally increases with elapsed time and cached (i.e.,

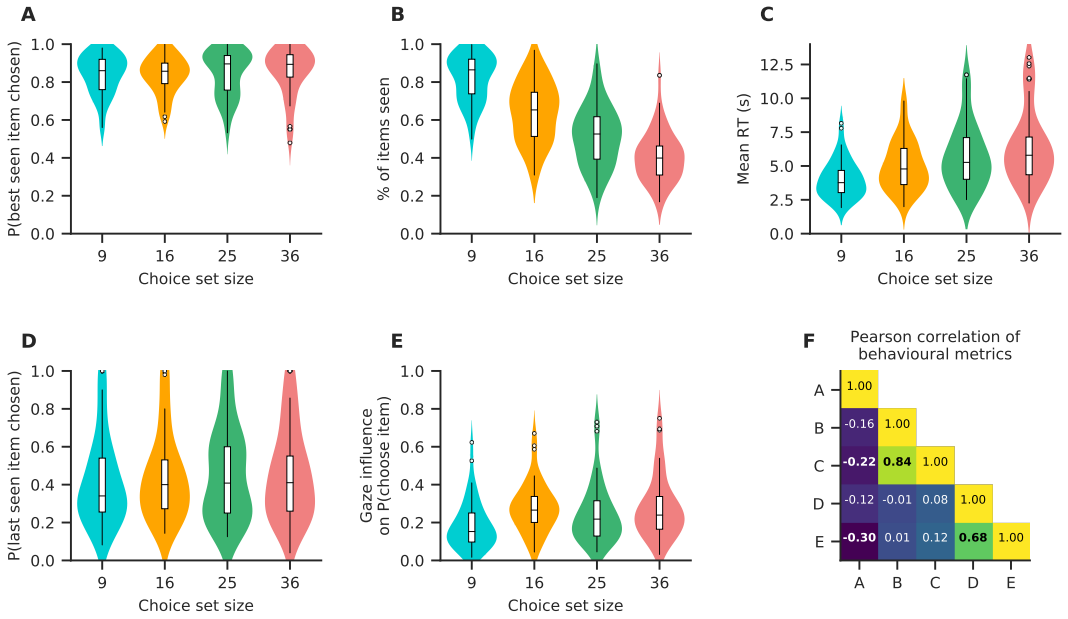

Figure 1.7: Choice psychometrics in MAFC. A: Subjects were very likely to choose one of the highest-rated items that they looked at in all choice set sizes. B-C: The fraction of items of a choice set that subjects looked at in a trial decreased with choice set size (B), while subjects' mean RTs increased (C). D: Subjects chose the item that they looked at last in a trial about half the time. E: Subjects generally exhibited a positive association of gaze allocation and choice behavior (as indicated by the gaze influence measure, describing the mean increase in choice probability for an item that is looked at longer than the others, after correcting for the influence of item value; see Appendix A.1). F: Associations of the behavioral measures shown in panels A - E (as indicated by Pearson's r correlation coefficients). Correlations with P-values smaller than 0.005 (Bonferroni corrected for multiple comparisons: 0.05/10) are printed in bold font. For a detailed view of the associations see Appendix Fig. A.7. Different colors represent different choice set size conditions. Violin plots show a kernel density estimate of the distribution of subject means with box plots inside of them.

the highest seen) value. We found that both had positive effects on subjects' stopping probability, in addition to a negative effect of choice set size (β = 2.67%, 95% HDI = [2.05, 3.28] per cached value, 2.26%, 95% HDI = [1.69, 2.80] per second, -0.22%, 95% HDI = [-0.24, -0.20] per item). Subjects' behavior was therefore qualitatively in line with the basic assumptions of the probabilistic satisficing model. Note that this finding does not allow us to discriminate between the probabilistic satisficing and evidence accumulation models, as they make very similar qualitative predictions about the relationship between stopping probability, time, and cached value.

Lastly, we tested individuals' association of gaze allocation and choice behavior (Fig. 1.3 B). All subjects exhibited positive values on the gaze influence measure in all choice set size conditions (with values ranging from 1.7% to 75%; β = 0.26%, 95% HDI = [0.15, 0.39] per item; Fig. 1.7 E; for further details on the gaze influence measure, see Appendix A.1), indicating an overall positive, but variable, association between gaze

allocation and choice. In general, subjects' probability of choosing an item increased with the item's gaze advantage (defined as the difference between the item's gaze and the maximum gaze of all other items in a choice set) and the item's relative rating (defined as the difference between the liking rating of an item and the mean rating of all other items in a choice set), while it decreased with the range of the ratings of the other items in a choice set and choice set size (β = 0.46%, 95% HDI = [0.4, 0.5] per percentage increase in gaze advantage, 3.6%, 95% HDI = [3.2, 4.0] per relative rating, -2.8%, 95% HDI = [-3.1, -2.4] per unit increase in the range of ratings of the other items, -0.16, 95% HDI = [-0.2, -0.1] per item).

To evaluate this finding in more detail, we probed three distinct predictions of the framework of gaze-driven evidence accumulation:

According to the framework of gaze-driven evidence accumulation, subjects who exhibit a stronger association of gaze and choice should generally also exhibit a lower probability of choosing the highest-rated item from a choice set (as has been shown for smaller choice sets; Fig. 1.3 F and Fig. 1.6 C). For these subjects, gaze allocation can bias the decision process towards items that have a lower value, if these are looked at longer. In line with this prediction, the gaze influence measure was negatively correlated with individuals' probability of choosing the highest-rated item that they have seen in a trial, such that a lower probability of choosing the highest-rated item was associated with a stronger association of gaze and choice (β = -0.036%, 95% HDI = [-0.053, -0.014] per percentage increase in probability of choosing the highest-rated seen item; Fig. 1.7 F).

Second, the framework of gaze-driven evidence accumulation predicts that subjects with a stronger association of gaze and choice should generally be more likely to look at the chosen item last in a trial, as evidence for the looked-at item is accumulated at a generally higher rate. In line with this prediction, subjects with a stronger association of gaze and choice were more generally likely to choose the item that they looked at last in a trial (β = 0.01%, 95% HDI =[0.008, 0.013] per percentage increase gaze influence; Fig. 1.7 F).

Lastly, the framework of gaze-driven evidence accumulation predicts that the probability that an individual chooses an item should increase with the duration of a single gaze to the item. In line with previous work (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011), we investigated this by studying the probability that subjects chose the first item that they looked at in a trial as a function of the duration of their first gaze to the item in the trial. Overall, the probability that subjects chose the item that they saw first in a trial increased as a function of the duration of their first gaze to the item (as well as the rating of the item) (β = 0.018%, 95% HDI = [0.014, 0.022] per ms, 6.0%, 95% HDI = [5.5, 6.6] per liking rating, -0.27%, 95% HDI = [-0.32, -0.22] per item).

Taken together, these findings demonstrate that subjects' choice behavior in MAFC does not match the assumptions of optimal choice or hard satisficing, while it qualitatively matches the assumptions of gaze-driven evidence accumulation and probabilistic satisficing. To further discriminate between the probabilistic satisficing model, GLAM, and independent evidence accumulation model (see section 1.3.2 and 1.3.4), we fitted these to each individual's choice and RT data in each set size condition (see Appendix A.1.5, A.1.6, and A.1.7 for details on the fitting procedures and Appendix Fig. A.8 for

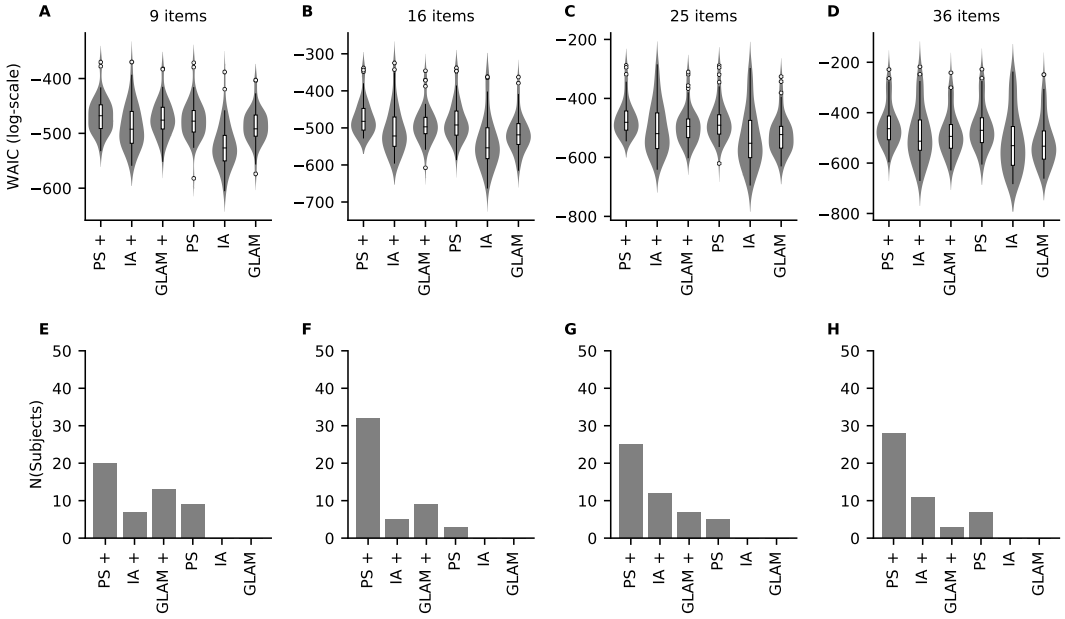

Figure 1.8: Individual relative model fit of the many-alternative choice models. A-D: Individual WAIC values for the probabilistic satisficing model (PS), GLAM, and independent evidence accumulation model (IA) in each choice set size condition. Model variants with an active account of gaze are marked with an additional "+". WAIC is shown on log-scale, such that larger values in WAIC generally indicate better model fit. Violin plots show a kernel density estimate of the distribution of WAIC values with box plots in side of them. E-H: Number of subjects in each choice set size condition best described by each model variant.

an overview of the parameter estimates).

In this comparison, we further considered two different accounts of gaze in the decision process. The passive account of gaze assumes that gaze allocation solely determines the set of items that are being considered; an item is only considered once it is looked at (to account for this in the GLAM, we excluded all items from the decision process that were not looked at in a trial). In contrast, the active account of gaze assumes that gaze drives the decision process by generating higher choice probabilities for items that are looked at longer. As recent research has indicated two distinct mechanisms through which gaze might influence the decision process, we accounted for both multiplicative (Krajbich et al., 2010, Krajbich and Rangel, 2011, Lopez-Persem et al., 2016, Smith and Krajbich, 2019, Tavares et al., 2017) and additive (Cavanagh et al., 2014, Westbrook et al., 2020) gaze bias effects in the modeling of the active account of gaze. Multiplicative effects discount the value of unattended items proportionally to their value, while additive effects add a constant boost to the value of the attended item (see section 1.3.2 and 1.3.4). Note that the model variants with a passive and active account of gaze were identical, except that we set $\gamma = 1$ and $\zeta = 0$ for the models with a passive account of gaze, thereby effectively removing the gaze bias mechanism from these models.

Overall, the response behavior of the vast majority of subjects was generally best

captured by the model variants with an active account of gaze, according to the WAIC (Vehtari et al., 2017) (82% (40/49), 94% (46/49), 90% (44/49), and 86% (42/49) subjects for choice sets with 9, 16, 25, and 36 items). Specifically, the choice behavior of the majority of subjects in all set size conditions was best described by the probabilistic satisficing model with an active account of gaze (41% (20/49), 65% (32/49), 51% (25/49), and 57% (28/49) in choice sets with 9, 16, 25, and 36 items respectively; Fig. 1.8). In the choice sets with 9 and 16 items, the GLAM variant with an active account of gaze best matched the response behavior of the largest group of the remaining subjects (27% (13/49) and 18% (9/49) subjects respectively; Fig. 1.8 A-B, E-F), while in the choice sets with 25 and 36 items the independent evidence accumulation model variant with an active account of gaze best matched the response behavior of the largest group of the remaining subjects (24% (12/49) and 22% (11/49) subjects respectively; Fig. 1.8 C-D, G-H).

Similar to our first empirical study (see Appendix Fig. A.5), we generally found evidence for both additive and multiplicative gaze bias effects in the GLAM and probabilistic satisficing model, as individual γ and ζ estimates were widely distributed in all choice set sizes (Appendix Fig. A.8 A-B, I-J). Interestingly, the independent evidence accumulation model did not exhibit any evidence for additive gaze bias effects (as individual ζ estimates were generally close to 0; Appendix Fig. A.8 R), but instead stronger multiplicative gaze biases (as individual γ estimates were also generally close to 0; Appendix Fig. A.8 Q).

To also probe the ability of the probabilistic satisficing model, GLAM, and independent evidence accumulation model to capture individuals' choice behavior on an absolute level, we simulated choice and RT data for each subject with each fitted model (Fig. 1.9; see Appendix A.1.8 for details on the simulation procedure). We restricted this analysis to the models with an active account of gaze as they best captured the response behavior of the vast majority of subjects. To assess the fit of the simulated data, we performed the following regression analysis: For each model and each behavioral measure, we regressed the measure onto a binary variable, indicating whether each value on this measure comes from the empirically observed data or from the model simulations. If the regression estimate of the indicator variable differed from 0, the model predictions deviate meaningfully from observed data. The probabilistic satisficing model and the GLAM both accurately captured subjects' mean RT (Fig. 1.9 A-B; β = -43 ms, 95% HDI = [-471, 421] difference for the probabilistic satisficing model; β = -321 ms, 95% HDI = [-729, 108] difference for the GLAM), while slightly underestimating subjects' probability of choosing the highest-rated item from a choice set (Fig. 1.9 D-E; β = -14.43%, 95% HDI = [-17, -11.67] difference for the probabilistic satisficing model; β = -13.04% ms, 95% HDI = [-15.71, -10.07] difference for the GLAM). The independent evidence accumulation, on the other hand, underestimated subjects' mean RT (Fig. 1.9 C; β = -1140 ms, 95% HDI = [-1522, -744] difference) and strongly underestimated subjects' probability of choosing the highest-rated item from a choice set (Fig. 1.9 F; β = -42.97%, 95% HDI = [-45.73, -40.31] difference). In contrast, only the GLAM and independent evidence accumulation model accurately captured subjects' gaze influence on choice probability (Fig. 1.9 H-I; β = -2.55%, 95% HDI = [-5.21, 0.17] difference for the GLAM; β = -1.29%, 95% HDI = [-3.64, 1.12] difference for the independent evidence

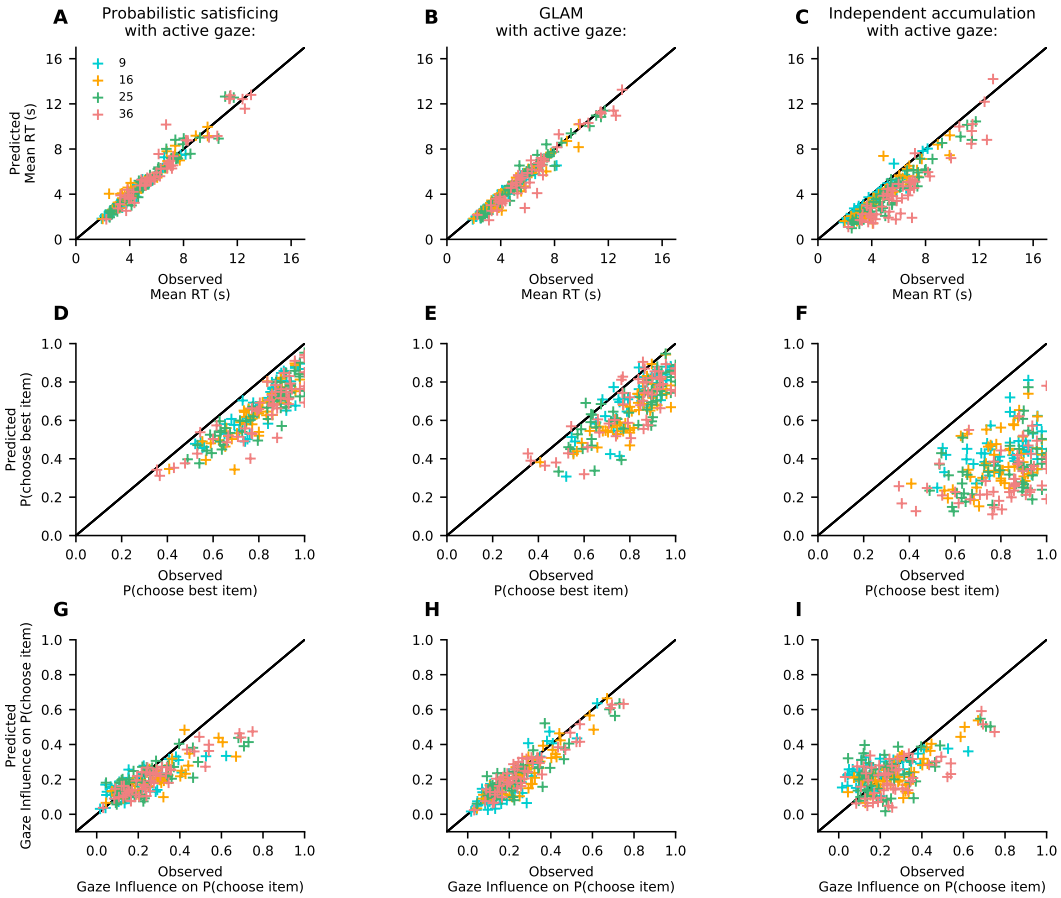

Figure 1.9: Individual predictions of behavioral metrics by the model variants with an active account of gaze in the decision process for the many-alternative choice dataset. A-C: The probabilistic satisficing model and GLAM accurately capture subjects' mean RT, which the independent evidence accumulation model slightly underestimates. D-F: All models underestimate subjects' probability of choosing the highest-rated item that they have seen in a trial. G-I: The GLAM and independent evidence accumulation model both accurately capture subjects' gaze influence, which the probabilistic satisficing model underestimates. See the main text for the corresponding statistical analyses. Model predictions are simulated using parameter estimates obtained from individual model fits (for more details on the simulation and fitting procedures, see Appendix A.1.5, A.1.6, A.1.7, and A.1.8). Crosses indicate individual participants, while colors indicate choice set size conditions.

accumulation model), which the probabilistic satisficing model slightly underestimated (Fig. 1.9 G; β = -4.94%, 95% HDI = [-7.26, -2.48] difference).

Taken together, these findings indicate an active role of gaze allocation in MAFC, similar to findings in smaller choice sets (Cavanagh et al., 2014, Krajbich et al., 2010, Krajbich and Rangel, 2011, Tavares et al., 2017). Subjects' response behavior qualitatively did not match the assumptions of optimal choice or hard satisficing, while subjects exhibited an overall positive association of gaze allocation and choice probability. Fur-

ther, the choice behavior of the vast majority of subjects in all set size conditions was best described by models that include an active account of gaze in the decision process, when compared to model variants with a passive account of gaze. In contrast to findings in smaller choice sets, the choice behavior of the majority of subjects in all choice set sizes best matched a probabilistic version of satisficing choice, in which the value of the seen items is increased through gaze allocation.

Visual search

To establish a general understanding of the visual search process in MAFC, we also performed an exploratory analysis of participants' visual search behavior.

The probability that participants looked at an item in a choice set increased with the item's liking rating, while decreasing with choice set size (β = 2.9%, 95% HDI = [1.6, 2.3] per liking rating, -1.4%, 95% HDI = [-1.5, -1.3] per item; Fig. 1.10 A-D). The probability that participants' gaze returned to an item also increased with the item's liking rating, while decreasing with choice set size (β = 1.6%, 95% HDI = [1.4, 1.8] per liking rating, -0.65%, 95% HDI = [-0.74, -0.55] per item; Fig. 1.10 A-D).

Similarly, the duration of a gaze towards an item increased with the item's liking rating, while decreasing with set size, for the initial gaze to an item (β = 9 ms, 95% HDI = [7, 10] per liking rating, -0.7 ms, 95% HDI = [-0.9, -0.5] per item; Fig. 1.10 E-H), as well as all subsequent gazes to the same item (β = 15 ms, 95% HDI = [10, 19] per liking rating, -2.2 ms, 95% HDI = [-2.5, -1.8] per item; Fig. 1.10 E-H).

On average, the initial gaze to an item was shorter than all later gazes to the same item in the same trial (β = 42 ms, 95% HDI = [35, 49] increase from initial to returning gazes; Fig. 1.10 I-L), consistent with prior findings (Krajbich et al., 2010). The duration of the last gaze in a trial was dependent on whether it was to the chosen item or not (Fig. 1.10 I-L): last gazes to the chosen item were in general longer in duration than middle gazes (excluding the first and last gaze of a trial) (β = 88 ms, 95% HDI = [50, 126] difference between last gazes to the chosen item and middle trial gazes), while last gazes to non-chosen items were generally shorter than middle trial gazes (β = -72 ms, 95% HDI = [-82, -62] difference between last gazes to non-chosen items and middle trial gazes).

We also studied subjects' visual search trajectories (Fig. 1.11): For each trial, we first normalized time to a range from 0 - 100% and then binned it into 10 % intervals. We then extracted the liking rating, position and size for each item in a trial (see Appendix A.1.3). An item's position was encoded by its column and row indices in the square grid (see Fig. 1.1 E; with indices increasing from left to right and top to bottom). All item attributes were centered with respect to their trial mean in the choice set (e.g., a centered row index of -1 represents the row one above the center, whereas a centered item rating of -1 represents a rating that is one below the average of all item ratings in that choice set). For each normalized time bin, we computed a mixed effects logit regression model (see Appendix A.1.4), regressing the probability that an item was looked at onto its attributes.

In general, subjects began their search at the center of the screen (Fig. 1.11 A-B; as indicated by regression coefficients close to 0 for the items' row and column positions

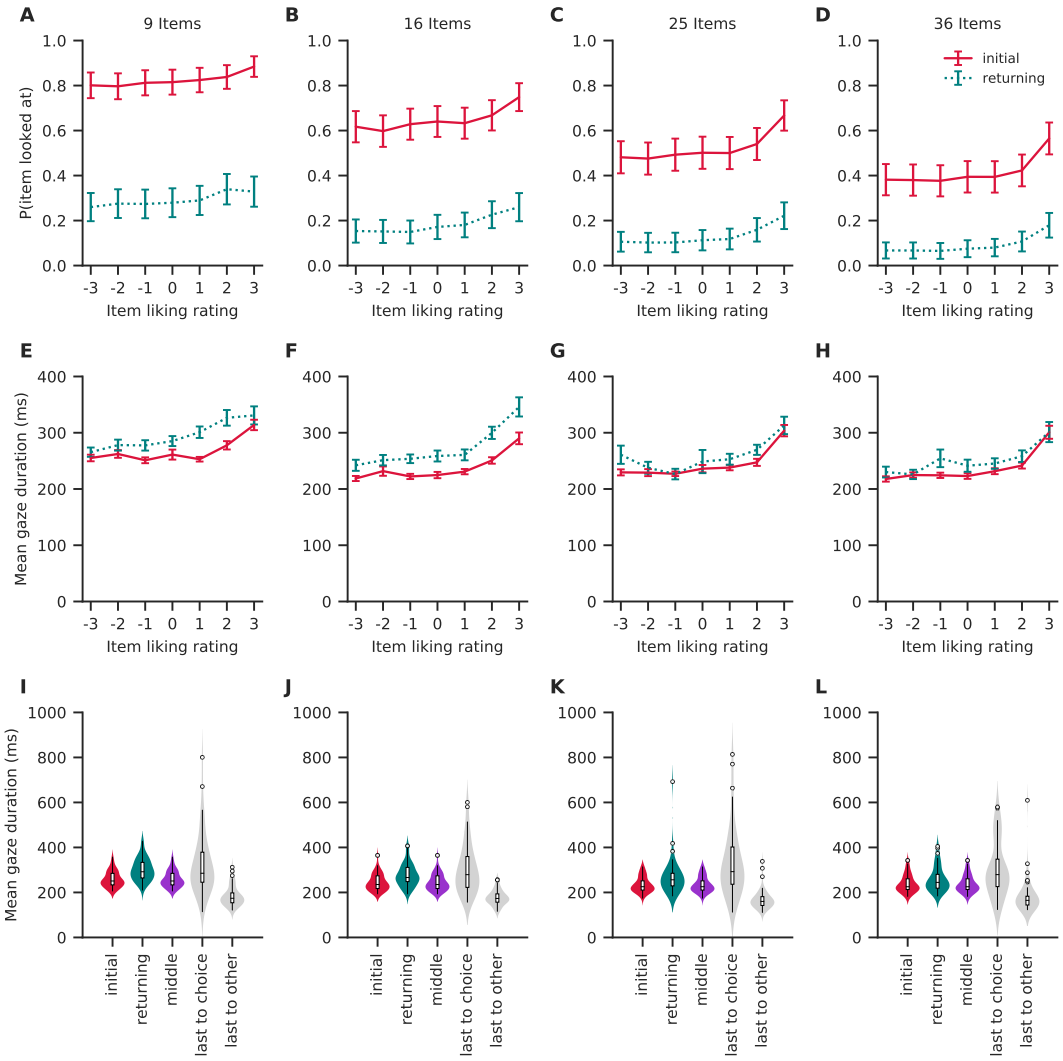

Figure 1.10: Gaze psychometrics in MAFC. A-D: The probability of looking at an item increases with its liking rating. E-H: The mean duration of item gazes increases with the liking rating of the item. I-L: Initial gazes to an item are in general shorter in duration than all subsequent gazes to the same item in a trial. The last gaze of a trial is in general shorter than all middle gazes (defined as all trial gazes except for the first and last) when it is not to the chosen item. It is, however, in general longer than all middle gazes when it is to the chosen item. Red indicates all initial gazes to an item, green indicates all returning gazes to the same item within the same trial, purple indicates all middle gazes that are neither the first nor last of a trial, and gray indicates the last gaze of a trial. Bar heights and colored lines indicate mean values with standard errors. Violin plots show a kernel density estimate of the distribution of subject means with box plots inside of them.

in the beginning of a trial), coinciding with the preceding fixation cross (Fig. 1.1 E). Subjects then typically transitioned to the top left corner (Fig. 1.11 A-B; as indicated by increasingly negative regression coefficients for the items' row and column positions in

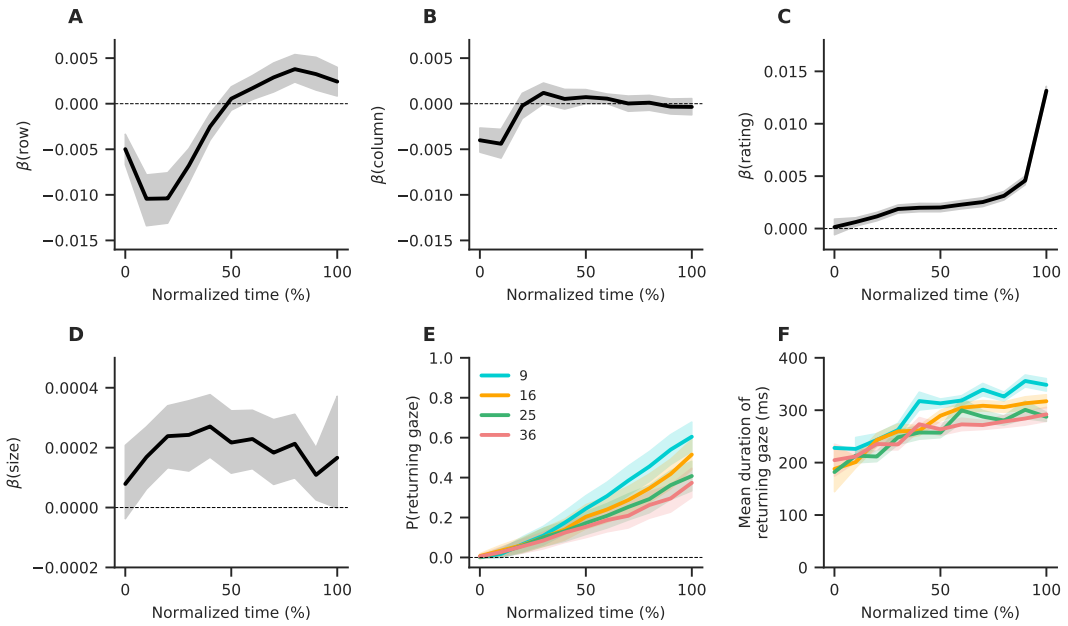

Figure 1.11: Visual search trajectory in MAFC: A-D: Black lines represent the fixed effects coefficient estimates (with 95% HDI intervals surrounding them) of a mixed effects logit regression analysis for each normalized trial time bin, regressing the probability that an item was looked at onto its centered attributes (row (A) and column (B) position, liking rating (C) and size (D)). Subjects generally started their search in the center of the choice screen, coinciding with the preceding fixation cross, and then transitioned to the top left corner (as indicated by decreasing regression coefficients for the items' row (A) and column positions (B)). From there, subjects generally searched from top to bottom (as indicated by slowly increasing regression coefficients for the items' row positions (A)), while also focusing more on items with a high liking rating (C) and a larger size (D). E-F: Over the course of a trial, the probability that subjects returned with their gaze to an item increased steadily (β = 10%, 95% HDI = [9, 11] per second, -0.73%, 95% HDI = [-0.80, -0.66] per item), as did the durations of these gazes (β = 14 ms, 95% HDI = [12, 17] per second, -2.9 ms, 95% HDI = [-3.3 -2.5] per item). Colored lines indicate mean values for each set size condition with standard errors surrounding them.

the beginning of a trial) and then moved from top to bottom (Fig. 1.11 A; as indicated by the then increasingly positive regression coefficients for the items' row positions). Over the course of the trial, subjects generally focused their search more on highly-rated (Fig. 1.11 C) and larger items (Fig. 1.11 D), while the probability that their gaze returned to an item, as well as the duration of returning gazes, steadily increased (Fig. 1.11 E-F). Overall, the fraction of total trial time that subjects looked at an item was dependent on all four item attributes as well as the number of items contained in the choice set (β = 0.5%, 95% HDI = [0.44, 0.58] per liking rating, 0.02%, 95% HDI = [0.008, 0.02] per percentage increase in size, -0.19%, 95% HDI = [-0.24, -0.15] per row position, -0.044 , 95% HDI =[-0.075, -0.007] per column position, -0.177, 95% HDI = [-0.18, -0.174] per item). Yet, the effects of item position and size on the search process decreased over time (Fig. 1.11 A-B, D).

We also tested whether these item attributes influenced subjects' choice behaviour. However, the probability of choosing an item did not depend on the size or position of the item, but was solely dependent on the item's rating and the choice set size (β = 4.0%, 95% HDI = [3.6, 4.4] per rating, 0.02%, 95% HDI = [-0.02, 0.06] per percentage increase in item size, -0.03%, 95% HDI = [-0.1, 0.04] per row, -0.01%, 95% HDI = [-0.07, 0.06] per column, -0.24%, 95% HDI = [-0.25, -0.23] per item).

To also better understand the relationship of individuals' visual search and choice behavior, we studied the association between the influence of an items' size, rating, and position on individuals' gaze allocation and the metrics of individuals' choice behavior reported in Figure 1.7 (namely, individuals' mean RT, fraction of looked-at items in a choice set, probability of choosing the highest-rated item from the set of seen items, and gaze influence on choice probability) (for an overview, see Appendix Fig. A.9). To quantify the influence of the item attributes on gaze allocation, we computed a mixed effects regression model for each individual in the data, regressing gaze (defined as the fraction of trial time that the individual looked at an item; scaled 0 - 100 %) onto the four item attributes (row, column, size, and rating) and choice set size. Individuals with a strong influence of item rating on gaze allocation generally looked at fewer items of a choice set (β = -17%, 95% HPI = [-32, -3] per unit increase in $\beta(rating)$; Appendix Fig. A.9 H), exhibited a higher probability of choosing the highest-rated item that they have seen in a trial (β = 14%, 95% HDI = [5, 24] per unit increase in $\beta(rating)$; Appendix Fig. A.9 P), and had a higher probability of choosing the last seen item (β = 40%, 95% HDI = [18, 62] per unit increase in $\beta(rating)$; Appendix Fig. A.9 T). Similarly, individuals' who exhibited a strong influence of item size on gaze allocation also generally looked at fewer items of a choice set (β = -115%, 95% HDI = [-211, -11] per unit increase in $\beta(size)$; Appendix Fig. A.9 G), exhibited shorter mean RTs (β = -18 s, 95% HDI = [-33, -3] per unit increase in $\beta(size)$; Appendix Fig. A.9 K), and were less likely to choose the item that they looked at last in a trial (β = -199%, 95% HDI = [-366, -26] per unit increase in $\beta(size)$; Appendix Fig. A.9 S). Lastly, subjects with a strong influence of item column on gaze allocation also generally exhibited longer mean RTs (β = 3.95 s, 95% HDI = [0.29, 8.05] per unit increase in $\beta(column)$; Appendix Fig. A.9 J). We did not find any other statistically meaningful association of the visual search and choice metrics (for an overview, see Appendix Fig. A.9).

Taken together, these findings indicate that individuals' visual search in MAFC is dependent on the value of the items in a choice set: The probability that an item was looked at, as well as the duration of a gaze to this item, increased with the item's liking rating. This trend also increased over the course of a trial, while the influence of other determinants of visual search, such as the the items' positions and size, decreased. Individuals with a strong influence of value on gaze allocation also generally looked at fewer items of a choice set, while being more likely to choose the highest-rated and last seen items of a trial. Interestingly, the strength of individuals' gaze allocation on choice probability (as indicated by the gaze influence measure; see Appendix A.1) was not meaningfully associated with any of the measures of individuals' search behavior (Appendix Fig. A.9 U-X), indicating that it cannot be explained by a specific type of visual search behaviour.

1.5 Conclusion

In this work, we investigated the association of gaze allocation and choice behavior across five datasets, spanning 167 individuals, six choice set sizes (2, 3, 9, 16, 25, and 36), and two choice domains (value-based and perceptual).

A key contribution of this work is the introduction of the gaze-weighted linear accumulator model (GLAM; see section 1.3.2), which allowed us to study the association of gaze allocation and choice behavior on the level of the individual and and in choice situations with many alternatives. The GLAM assumes that individuals accumulate evidence in favor of each choice alternative and make a choice as soon as the cumulative evidence for one alternative reaches a choice threshold. Importantly, the accumulation process is biased by gaze allocation, with generally discounted accumulation rates for unattended items.

In the following, we will summarize the key findings of this work for each of our research hypotheses (see section 1.2).

1. **Individuals exhibit a gaze bias in simple choice situations with few alternatives.** To test this hypothesis, we analyzed four previously published choice datasets, spanning 118 individuals, two choice set sizes (two and three items), and two choice domains (value-based and perceptual choice). We fitted the GLAM to data of each individual and compared its fit to a model variant without any gaze bias. The GLAM variant with gaze bias generally captured the RT and choice data of most individuals better than the model variant without gaze bias; in a likelihood-based model comparison and in an out-of-sample prediction. Our findings thereby confirm this hypothesis.

2. **The strength of the gaze bias varies between individuals.** To test this hypothesis, we performed two analyses: First, we studied the behavioral association of gaze allocation and choice probability for the 118 individuals in the four simple choice datasets of our first study (Fig. 1.3 B). Subsequently, we studied the gaze bias parameter estimates from the individual GLAM model fits to these datasets (see Appendix Fig. A.1). In both analyses, individuals generally exhibited a positive association of gaze allocation and choice (with generally higher choice probabilities for items that were looked at longer), while the strength of this association was highly variable between individuals. Our findings thereby confirm this hypothesis.

3. **Individuals with a strong gaze bias are generally less likely to choose the best item from a choice set compared to individuals with a weak gaze bias.** To test this hypothesis, we studied the association of individuals' probability of choosing the best item from a choice set with a behavioral measure of gaze influence (Fig. 1.3 F), and the gaze bias parameter estimates from the individual GLAM model fits of study I (Fig. 1.6 C). Both were correlated with individuals' probability of choosing the best item, such that a stronger association of gaze and choice was generally associated with a lower probability of choosing the best item from a choice set. Our findings thereby confirm this hypothesis.

4. **Gaze allocation and choice behavior exhibit a similar positive association in many-alternative choices as they do in smaller choice sets.** To test this hypothesis, we analyzed the choice behavior of 49 individuals, who made repeated value-based choices from set of 9, 16, 25, and 36 alternatives (for an overview of the Thomas 2020 dataset, see section 1.3.1 and Thomas et al., 2020). As in smaller choice sets, all individuals exhibited an overall positive association of gaze allocation and choice probability in all choice set sizes, while the strength of this association was highly variable between individuals (Fig. 1.7 E). Individuals' response behavior did not match the assumptions of optimal choice and hard satisficing, as individuals did not look at every item of a choice set and chose the item they looked at last only half the time. In general, individuals' response behavior was best captured by models that included a gaze bias mechanism, while the response behavior of the majority of individuals in all choice set sizes was best described by a probabilistic satisficing model with an active account of gaze (in which gaze allocation increases the values of the items). Our findings thereby confirm this hypothesis.

We further performed an exploratory analysis of individuals' visual search behavior in MAFC. Individuals were generally more likely to look at items that have a higher value and to look at these items longer. In general, individuals started their visual search in the top left corner of a choice screen and then worked their way from left to right and top to bottom, while also focusing more on larger items. Over the course of a trial, the influence of item position and size on individuals' visual search decreased, while individuals focused their search more on the highly valued items of a choice set.

In summary, this work has demonstrated that individuals exhibit a positive association of gaze allocation and choice behavior in simple choice situations involving two to 36 items. Importantly, the strength of this association is highly variable between individuals, such that accounting for this variability is necessary to accurately explain and predict individuals' choice behavior.

1.6 Discussion

1.6.1 The gaze-weighted linear accumulator model

The GLAM is inspired by the multi-alternative aDDM (Krajbich and Rangel, 2011) and can be used to study gaze biases at the level of the individual and in choice situations with many alternatives. It assumes that individuals accumulate evidence in favor of each available item and make a choice as soon as the cumulative evidence for one item reaches a choice threshold. Importantly, the accumulation process is biased by gaze behavior, with discounted accumulation rates for unattended items. The model is statistically and computationally tractable, making it readily extendable to novel choice tasks and research questions. Generally, the GLAM can be seen as a way to sidestep the complex problem of simulating individual fixation trajectories, as required by the aDDM. The GLAM solely uses the observed distribution of gaze to the items over the course of the trial. In contrast, the aDDM is fitted to empirical data using model simulations, which rely on an accurate simulation of the fixation trajectories. Although researchers have started to explore generative fixation models in simple decision-making tasks (e.g., Callaway et al., 2019, Jang et al., 2020, Towal et al., 2013), this is often not feasible or not of main interest to researchers trying to understand the influence of gaze allocation on choice behavior. Here, the GLAM provides a tractable, but simplified, alternative to the aDDM that solely requires trial-level statistics, namely, overall gaze proportions (next to the RTs, choices and item values). As a side effect, this allows the application of the GLAM to situations in which only limited trial-level data are available (for example, the Folke 2016 dataset, which only contains trial-averaged gaze data). In theory, a similar simplification of the multialternative aDDM (Krajbich and Rangel, 2011) would be possible, but would result in a model highly similar to the GLAM. Furthermore, fitting such a simplified aDDM variant would still rely on simulations of the decision-making process as no analytical solution exists for race models with negative drift terms. These simulations are particularly costly in the case of the aDDM, whereby every trial represents a unique condition due to the incorporation of trial-specific eye movement data.

1.6.2 Limitations of the gaze-weighted linear accumulator model

The GLAM's usage of trial-level gaze data makes it computationally tractable. Yet, due to this simplification, the GLAM does not capture dynamics of the decision process on the level of single fixations. It does, for example, not differentiate in which sequence the available choice alternatives were looked at. If these fine-grained dynamics are of interest to the researcher, the multi-alternative aDDM (Krajbich and Rangel, 2011) can be used instead, as it does not average-out the fixation-dependent changes in evidence accumulation rates throughout the trial. Keeping this level of detail in the aDDM, however, comes at the computational cost of extensive model simulations and the need to build a generative fixation model.

While we have demonstrated that the GLAM captures individuals' gaze bias and choice behavior well in choice situations involving only two alternatives, there exist other computational approaches that can estimate the gaze bias of an individual in

binary decisions: If response times are of interest to the researcher, the gaze bias can be estimated in the form of a gaze-weighted DDM (e.g., Cavanagh et al., 2014, Lopez-Persem et al., 2016). Similar to the GLAM, this approach also aggregates over the dynamics of fixation process within a trial by utilizing the fraction of trial time that each item was looked at. In contrast to the GLAM, however, gaze-weighted DDM approaches describe the decision process as a single accumulator that evolves between two decision bounds (each representing one of the two decision alternatives). For two-alternative choices, the DDM is generally viewed as optimal in the sense of the sequential probability ratio test (Ratcliff et al., 2016). For two-alternative choice scenarios, in which response times are not of interest to the researcher, Smith and colleagues (Smith et al., 2019) proposed a method for estimating the aDDM gaze bias parameter through a random utility model.

1.6.3 Individual differences in the association of gaze allocation and choice behavior

Our analyses confirmed the need to account for individual variability in the gaze bias, as we found substantial variability across individuals in the influence of gaze on choice in all choice datasets that we analyzed, spanning six different choice set sizes (2, 3, 9, 16, 25, and 36 items) and two choice domains (value-based and perceptual choice). This interindividual variability was previously hidden in group-level analyses (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011).

Given that the influence of gaze on choice is variable among individuals, a single gaze bias parameter for the whole group would not fit all individuals well, and therefore result in inferior predictive performance of the model. On the one hand, individuals who exhibit a weaker link between gaze allocation and choice behavior than the group average would falsely be predicted to make choices less consistent with item values, and driven more by looking behavior. On the other hand, predictions for individuals' choices with a stronger link than the group average would not contain enough influence of gaze. Accounting for individual differences in the link between gaze allocation and choice behavior opens important avenues for future research that focus on the specific determinants of these differences. For example, are these differences best characterized as a trait (stable within a person, but variable between persons), a state (variable within a person, between different situations or contexts) or both (variable between persons and contexts)?

1.6.4 The role of gaze allocation in choices from many alternatives

As in prior work, our findings firmly reject a model of complete search and maximization in MAFC (e.g., Caplin et al., 2011, Pieters and Warlop, 1999, Reutskaja et al., 2011, Simon, 1959, Stüttgen et al., 2012): Subjects do not look at every item and they do not always choose the best item that they have seen. Our data also clearly reject a hard satisficing model: subjects choose the last item they look at only half of the time. Additionally, we find that subjects' choices are strongly dependent on the actual time that they spend looking at each alternative and can thereby not be fully explained by simply accounting for the set of examined items, as a passive account of gaze in

the decision process would assume. This stands in stark contrast to many models of consumer search and rational inattention (e.g., Caplin et al., 2011, Masatlioglu et al., 2012, Matějka and McKay, 2015, Sims, 2003), which ascribe a more passive role to visual attention, by viewing it as a filter that creates consideration sets (by attending only to a subset of the available alternatives) from which the decision maker then chooses.

To better understand the computational mechanism underlying individuals' choices, we compared the fit of a probabilistic satisficing model (based on findings by Reutskaja et al., 2011), the GLAM, and an independent evidence accumulation model to each individuals' choice and RT data in choice sets with 9, 16, 25, and 36 items. We additionally tested two versions of each model, including an active or passive account of gaze in the decision process. In general, the behavior of the majority of individuals in all choice set sizes was best described by the model variants that included an active account of gaze in the decision process, while the probabilistic satisficing model with gaze bias provided the overall best fit in all choice set sizes. These findings therefore bridge the divide between research in smaller choice sets, arguing for processes of gaze-driven evidence accumulation (e.g., Krajbich et al., 2010, Krajbich and Rangel, 2011, Lopez-Persem et al., 2016, Smith and Krajbich, 2019, Tavares et al., 2017), and traditional models of many-alternative choice (e.g., Reutskaja et al., 2011).

Overall, our findings indicate that attention takes up an active role in MAFC, similar to choices from smaller choice sets, by guiding preference formation (e.g., Armel et al., 2008, Gluth et al., 2018, Krajbich et al., 2010, Krajbich and Rangel, 2011, Smith and Krajbich, 2018).

1.6.5 Visual search in large choice sets

We also conducted an exploratory analysis to better understand subjects' search patterns in MAFC. In general, subjects started their search in the top left corner of the choice set, then gradually worked their way from top to bottom, while also focusing more on bigger items. In contrast to the binary choice setting, we found that the probability that an item was looked at, as well as the duration of a gaze to this item, increased with the item's rating, and that this trend also increased over the course of a trial. In a sense this is not surprising: When deciding between two alternatives, you are merely trying to compare one to the other. In that case attending to either alternative is equally useful in reaching the correct decision. However, with many choice alternatives, it is in your best interest to quickly identify the best alternatives in the choice set and exclude all other alternatives from further consideration (e.g., Hauser and Wernerfelt, 1990, Payne, 1976, Reutskaja et al., 2011, Roberts and Lattin, 1991). Given this search and decision process we might expect that subjects' choices are more driven by their gaze in the later stages of the decision, when they focus more on the highly rated items in the choice set, than in the earlier stages of the search, when subjects' gaze is driven by the items' positions and sizes. Indeed, we found that only the items' ratings predicted subjects' choice behavior, not their positions or sizes.

Chapter 2

The analysis of fMRI data through deep learning models

This chapter covers the contributions from Thomas et al. (2019a) and Thomas et al. (2019c). I would like to thank my co-authors for allowing me to use text and figures from these publications.

2.1 Introduction

2.1.1 Challenges in the analysis of functional magnetic resonance imaging data

A central goal of functional neuroimaging research is to understand the association between the cognitive state of an individual (e.g., while deciding whether a group of dots is moving to the left or right, or whether to accept or reject a risky gamble) and the underlying brain activity (Gazzaniga, 2006). Understanding this association is challenging, as the brain constantly performs many computations in parallel and the activity associated with a specific cognitive state can be spatially widely distributed. Untangling a specific cognitive state from all other occurring computations therefore requires well-thought-out experiment paradigms and appropriate statistical and computational methods.

While many of the findings presented in this work can be generalized to other types of functional neuroimaging data, this work is focused on functional Magnetic Resonance Imaging (fMRI; Huettel et al., 2004) data. FMRI measures local changes in blood flow that occur in response to the increased oxygen demand of active neurons (the haemodynamic response; Lindquist et al., 2009). The haemodynamic response results in a change in the local ratio of oxyhemoglobin and deoxyhemoglobin, which can be detected by fMRI through their different magnetic properties (the blood-oxygen-level-dependent, or BOLD signal). While fMRI has a comparably high spatial resolution, it generally suffers from a low temporal resolution (for a detailed discussion of this issue, see Murphy et al., 2007). FMRI datasets are thereby notoriously high-dimensional (with many hundred thousand dimensions (or voxels) for each collected fMRI volume), while only containing a few hundred volumes for each individual in a given dataset. This setting of high-dimension and low-sample size data poses one of the greatest challenges for the analysis of fMRI data, as analyses performed on these datasets generally lack statistical power (for a detailed discussion on this issue, see Cremers et al., 2017, Marek et al., 2020).

To deal with the high dimensionality and low sample size of conventional fMRI datasets, many prominent analysis approaches include limiting assumptions by either analyzing the data of single voxels (or groups of voxels) independent from one another, utilizing simple linear mappings between cognitive states and brain activity, or focussing solely on the group-level. While these limitations are often well justified and have enabled us to build an understanding of the mapping between many cognitive states and brain activity, there is a growing demand for computational methods that can better capture the non-linear temporal and spatial dependencies of whole-brain activity as well as their intra- and inter-individual variability (e.g., Loula et al., 2018, Mensch et al., 2018, Morioka et al., 2020, Rosa et al., 2015, Varoquaux and Craddock, 2013, Varoquaux et al., 2011).

2.1.2 The promise and challenges of deep learning for functional neuroimaging research

DL methods can generally be described as a class of representation-learning methods, with multiple levels of abstraction (Goodfellow et al., 2016, LeCun et al., 2015). At each level, the representation of the input data is transformed by a simple but non-linear function. The resulting hierarchical structure of non-linear transforms enables DL methods to learn complex functions and to identify intricate signals in noisy data, by projecting the input data into a higher-level representation in which those aspects of the input data that are irrelevant for identifying a target signal are suppressed and those aspects that are relevant are amplified. With this higher-level perspective, DL methods can associate a target signal with variable patterns in the input data. In general, DL methods can autonomously learn these representation from sufficiently large datasets and therefore do not require a thorough prior understanding of the mapping between input data and analysis target. These characteristics have made DL methods widely popular in many research and industry applications, where they have demonstrated strong empirical success compared to many conventional machine learning (ML) approaches (for an overview, see LeCun et al., 2015). DL methods thereby seem ideally suited for the analysis of fMRI data, where intricate, highly variable patterns of brain activity are hidden in high-dimensional datasets and the mapping between cognitive state and brain activity is often unknown.

Neuroimaging research has further begun collecting large corpora of experimental fMRI data, often comprising many hundred individuals (e.g., Alfaro-Almagro et al., 2018, Poldrack et al., 2013, Van Essen et al., 2013), thereby generally allowing for the application of DL models. With these datasets, researchers seek to gain deeper insights into the associations between the cognitive states of an individual and the underlying brain activity, while also studying the variability of these associations across the population.

In spite of the availability of these large datasets, two major challenges have so far prevented broad DL usage:

1. Due to the combination of many non-linear transforms, DL models generally act as *"black boxes"*, disguising any association between their input data and predictions. Even if a DL model is trained to accurately decode a set of cognitive states from

whole-brain fMRI data, it does not provide insight into the learned mapping between brain activity and cognitive states.

2. While several large and public fMRI datasets exist, conventional fMRI datasets suffer from high dimensionality and low sample size. A typical fMRI dataset contains a few hundred thousand voxels (or dimensions) for each fMRI volume, and only a few hundred volumes for each of up to a hundred individuals. In this setting, DL models (as well as more traditional ML approaches) are prone to overfitting (by too closely capturing those dynamics that are specific to a dataset), such that the trained models do not generalize well to new data (for a detailed discussion, see Claeskens and Hjort, 2008, Guthery et al., 2003).

The goal of this work is to investigate two solutions to these challenges, which utilize methods of explainable artificial intelligence and transfer learning.

2.1.3 Explaining the decoding decisions of deep learning models through relevance decomposition

The ability of DL methods to learn complex behaviors (for example, learning to drive autonomous cars or modelling quantum many-body systems; Chmiela et al., 2017, Schütt et al., 2017, 2018) has made them widely popular in many research and industry applications (for a detailed review, see Goodfellow et al., 2016, LeCun et al., 2015). This empirical success has led to a growing interest (and demand) to better understand the behavior of DL models and its association to the input data. To this end, scientists have advocated for the formation of a new field of research to study the behavior of intelligent artificial agents (for a detailed review, see Rahwan et al., 2019, Samek et al., 2019). One line of research within this field seeks to explain the predictions of DL models by relating these predictions to the features of the input data (for an overview, see Montavon et al., 2018, Samek et al., 2020, 2019). The resulting explanation techniques are thereby well-suited for the analysis of fMRI data, where researchers aim to identify the association between the cognitive states of an individual and the features of these data.

At a first level, explanation techniques can be divided into those that aim to obtain a global or local explanation. Global explanation techniques, such as activation maximization (Nguyen et al., 2016a,b), identify prototypical cases for a specific prediction (for instance, the image for which a classifier is most certain that it depicts a dog). Local predictions, on the other hand, relate the contribution of the individual features of an input to the prediction. While global and local explanation techniques can both be insightful for the application of DL methods to fMRI data, this work focuses on local explanation techniques because of their ability to analyze different levels of data granularity, from the level of individual time points to the level of subjects and groups of subjects. Specifically, we will focus here on the layer-wise relevance propagation (LRP; Bach et al., 2015, Montavon et al., 2017, Samek et al., 2020) technique, which has several desirable properties for the analysis of fMRI data.

The goal of the LRP technique is to identify the relevance (or contribution) of each dimension of an input sample to the prediction of a function. Conceptually, the

LRP technique builds on Deep Taylor Decomposition (DTD; as proposed by Montavon et al., 2017), which assumes that the function follows the structure of a DL model, with multiple subsequent layers that are each composed of multiple linear functions (or neurons) with added non-linearity. While the LRP technique generalizes to other function structures (see, for example, Kauffmann et al., 2020, Schnake et al., 2020), we will focus here on its application to DL models. To identify the contribution of each feature of an input to the prediction of a DL model (defined by the activation score in the output layer), the LRP technique redistributes the prediction in a backward propagation pass according to the contributions of each neuron in each layer to the prediction, until the input space is reached. Importantly, the redistribution process adheres to a conservation principle, such that the relevance that is attributed to a neuron is conserved in the relevances assigned to all neurons in the next lower layer to which this neuron is connected. For image data (such as those resulting from fMRI), these relevances can be visualized in the form of a heatmap in the input data space. This also allows the embedding of the resulting relevance data in existing fMRI analysis frameworks, as the relevance and fMRI data have the same dimensionality.

A recent comparison to other local explanation methods (for an overview, see Samek et al., 2020) demonstrated that LRP generally provides explanations that are faithful (such that they reliably and comprehensively identify the decision structure of a DL system; c.f. Samek et al. (2020)), interpretable for the human viewer, and have a low computational cost. The LRP technique thereby has several desirable properties for the analysis of fMRI data, which deals with high dimensional datasets and requires explanations that are are fast to compute, interpretable for the researcher, and representative of the decision structure of the analyzed DL model. The latter is especially relevant for functional neuroimaging research, as there often exists no ground truth for the association between cognitive state and brain activity, such that researchers must rely on the faithfulness of an explanation, in so far as it represents a reliable and comprehensive description of the learned mapping between brain data and cognitive states.

2.1.4 Improving the application of deep learning models to conventional fMRI datasets with transfer learning

The strong expressive power of DL models is directly related to their statistical and computational complexity, which results from the combination of many simple non-linear functions (for a general overview of DL methods, see Goodfellow et al., 2016, LeCun et al., 2015). Current state-of-the-art DL models can easily contain many million parameters (e.g., Devlin et al., 2019, He et al., 2015, Krizhevsky et al., 2012, Simonyan and Zisserman, 2015, Szegedy et al., 2015). This boundary has been pushed even further by the recently published GPT-3 natural language processing (NLP) model of OpenAI (Brown et al., 2020), which contains 175 *billion* parameters. Initial empirical evidence indicates that the ability of DL models to generalize between datasets and domains, as well as their general performance in the Turing test (which probes the ability of a machine to exhibit intelligent behaviour; Pinar Saygin et al., 2000), increases with the number of trainable parameters in a model (Brown et al., 2020). Yet, the vast and growing complexity of state-of-the-art DL architectures also represents one the biggest

challenges for their application to conventional datasets, in fields where data is scarce and data collection expensive.

For this reason, ML researchers have proposed many approaches to improve the performance of DL models in small datasets (e.g., Krogh and Hertz, 1992, Ng, 2004, Pascanu et al., 2013, Srivastava et al., 2014). One such approach which has had strong empirical success is the transfer learning method (for a detailed review, see Pan and Yang, 2010). The goal of transfer learning is to leverage the knowledge about a mapping between input data and target variable that can be learned from one dataset to subsequently improve the learning of a similar mapping in another dataset of a related domain. Transfer learning has been especially successful in computer vision and NLP, where large publicly available datasets exist (e.g., Bowman et al., 2015, Deng et al., 2009, Rajpurkar et al., 2016). Here, DL models are first *pre-trained* on these large datasets (e.g., to identify objects depicted in an image) and subsequently *fine-tuned* on smaller datasets of a related domain (e.g., to detect breast cancer in medical imaging; Khan et al., 2019). Pre-trained models generally exhibit faster learning and higher predictive accuracies, while also requiring generally less training data, when compared to models that are trained from scratch on the fine-tuning dataset (Yosinski et al., 2014).

There are two common types of fine-tuning for DL models. One is to hold the weights of the pre-trained model (or parts of the model) constant during fine-tuning (e.g., Rajaraman et al., 2018). The pre-trained model thereby acts as a feature extractor for a new model (or another part of the model) that is trained on its representations. The other common approach is to solely utilize the learned weights of the pre-trained model as a starting point for the fine-tuning (e.g., Samek et al., 2017, Thomas et al., 2019c). Conceptually, the pre-trained weights then act similar to a prior, such that the fine-tuned model has better chances to converge on a point in the parameter space that exhibits overall lower loss and better generalizability. Both approaches can also be combined in variable form, by freezing the weights of some parts of a pre-trained DL model, while allowing others to change (e.g., Guo et al., 2019, Hinterstoisser et al., 2018).

Over recent years, the field of functional neuroimaging has experienced a rise in the availability of large and public datasets, such as those provided by the Human Connectome Project (HCP; Van Essen et al., 2013), UK BioBank (Sudlow et al., 2015), National Institute of Health (e.g., the Adolescent Brain Cognitive Development (ABCD) Study; Casey et al., 2018), Max Planck Society (Babayan et al., 2019), and UCLA Consortium for Neuropsychiatric Phenomics (Poldrack et al., 2016). Further, there is a growing consensus among functional neuroimaging researchers to make their individual fMRI datasets available to the public through data sharing platforms like OpenNeuro (`https://openneuro.org/`). These developments have slowly paved the way for the field of functional neuroimaging to enter a big data era. The availability of these datasets has further raised the question whether transfer learning can be similarly beneficial for the application of DL models to fMRI, as it has been in other fields. First empirical evidence indicates that the transfer of knowledge (through DL models) between individuals and fMRI datasets is possible (e.g., Mensch et al., 2018, Zhang et al., 2018) and can improve the performance of DL models in smaller fMRI datasets (e.g., Mahmood et al., 2019, Thomas et al., 2019c, Zhang and Bellec, 2020).

2.2 Research questions and hypotheses

The application of DL models to whole-brain fMRI data has generally been hindered by two challenges: I) DL models act as "black boxes", disguising any relationship between a decoded cognitive state and the underlying brain activity; II) fMRI datasets are high-dimensional, while containing comparably few samples. In these high-dimension and low-sample size settings, DL models are at risk of overfitting.

The goal of this work is to investigate two solutions for these challenges, by the use of explanation techniques and transfer learning. Specifically, we investigate whether DL models, in combination with the LRP technique (Bach et al., 2015, Montavon et al., 2017, Samek et al., 2020), provide a means to accurately decode a set of cognitive states from whole-brain fMRI data and identify a biologically plausible association between the decoded cognitive state and brain activity. We further investigate whether transfer learning can improve the performance of DL methods in the high-dimension low-sample size setting of typical fMRI datasets.

A key contribution of this work is the introduction of the DeepLight framework (see section 2.3.3), which uses a DL model to decode a cognitive state from whole-brain fMRI data and subsequently relates the decoded cognitive state and brain activity, by interpreting the decoding decision with the layer-wise relevance propagation (LRP; Bach et al., 2015, Montavon et al., 2017) technique.

Specifically, this work tests the following set of hypotheses:

1. **A DL model can be trained to accurately decode a set of cognitive states from single whole-brain fMRI volumes.** Decoding approaches have been very successful in identifying cognitive states from fMRI data by the use of conventional ML classifiers (for an overview, see Haynes and Rees, 2006). DL methods have been shown to outperform many conventional ML approaches, in settings with high-dimensional and noisy data (for an overview, see LeCun et al., 2015). For this reason, we hypothesize that DL models can also be used to accurately decode cognitive states from single whole-brain fMRI volumes.

2. **The application of the LRP technique to the decoding decisions of a DL model that accurately decodes a set of cognitive states from whole-brain fMRI data allows to identify a biologically plausible association between the decoded cognitive states and brain activity.** Over recent years, the LRP technique has had strong empirical success in providing insights into the associations between input data and the predictions of DL models in many research fields (e.g., Hägele et al., 2020, Horst et al., 2019, Lapuschkin et al., 2019, Sturm et al., 2016). We therefore hypothesize that the application of the LRP technique to the decoding decisions of a DL model that accurately decodes a set of cognitive states from whole-brain fMRI data also provides meaningful insights into association between the decoded cognitive states and brain activity.

3. **A DL model that is pre-trained on a large fMRI dataset achieves overall higher decoding accuracies and requires less training time and data than a model variant with the same architecture that is trained from**

scratch, when both are applied to the fMRI data of an *independent experiment task*. ML research indicates that transfer learning, even between distant tasks, is generally beneficial for the application of DL models to conventional datasets, when compared to random weight initialization (Yosinski et al., 2014). For this reason, we hypothesize that transfer learning is similarly beneficial for the application of DL models to the fMRI data of different experiment tasks. Specifically, we hypothesize that a DL model which is pre-trained on the fMRI data of six out of the seven experiment tasks of the Human Connectome Project (HCP; Van Essen et al., 2013) outperforms a model variant that is trained from scratch, when both are applied to the fMRI data of the left-out seventh HCP experiment task.

4. **A DL model that is pre-trained on a large fMRI dataset achieves overall higher decoding accuracies and requires less training time than a model variant with the same architecture that is trained from scratch, when both are applied to the fMRI data of an *independent fMRI dataset*.** In line with our hypothesis 3, we further assume that transfer learning is beneficial for the application of DL models to the fMRI data of independent datasets. Specifically, we hypothesize that a DL model which is pre-trained on the fMRI data of the Human Connectome Project (Van Essen et al., 2013) outperforms a model variant that is trained from scratch, when both are applied to an fMRI dataset that is not part of the Human Connectome Project. To test this, we will use an fMRI dataset which was recently published by Nakai and Nishimoto (Nakai and Nishimoto, 2020). This dataset contains the data of six healthy human participants who repeatedly performed 103 simple naturalistic tasks.

5. **The DeepLight framework is generalizable to different DL model architectures.** Research in computer vision has demonstrated that two prominent DL model architectures gnerally perform well in learning the spatial dependency structure of sequential (or volumetric) data. One architecture separates the input data into a sequence of two-dimensional images, which are then processed by the combination of two-dimensional convolutional and recurrent DL elements (e.g., Donahue et al., 2015, Marban et al., 2019). The other directly applies three-dimensional convolutions to the input data, thereby accounting for all three spatial dimensions (e.g., Tran et al., 2015). We hypothesize that the DeepLight framework generalizes to both of these architectures, in that they achieve similar decoding performance and both learn a biologically plausible mapping between brain activity and cognitive states.

2.3 Methodology

2.3.1 Overview of datasets and experiment tasks

Human Connectome Project

Task	Cognitive states	Count	Duration (min)
WM	body, face, place, tool	4	5:01
Gambling	win, loss, neutral	3	3:12
Motor	left/right finger, left/right toe, tongue	5	3:34
Language	story, math	2	3:57
Social	interaction, no interaction	2	3:27
Relational	relational, matching	2	2:56
Emotion	fear, neutral	2	2:16
Total		**20**	**23:03**

Table 2.1: Overview of the HCP fMRI data. For each experiment task, the cognitive states, the number of cognitive states, and the duration of the task in each run are presented.

The task-fMRI data of the Human Connectome Project (Barch et al., 2013) includes seven tasks that were each performed in two separate runs (for a general overview, see Table 2.1; for details on the experiment paradigms, see Appendix B.1.1). For each task, participants were first provided with detailed instructions outside of the fMRI and only given a very brief reminder of the task and a refresher on the response button box mappings before the start of each task in the fMRI.

Working memory (WM): Participants were asked to decide in an N-back task whether a currently presented image (of body parts, faces, places or tools) is the same as a previously presented target image. The target image was either presented at the beginning of the experiment block (0-back) or participants were asked to decide whether the currently presented image is the same as the one presented two before (2-back).

Gambling: Participants were asked to guess whether the value of a card (with values between 1-9) is below or above 5. Participants won or lost if they guessed correctly/incorrectly. Trials were neutral if the value of the card was 5. The number on the card was dependent on whether the respective trial belonged to the reward, loss, or neutral task condition.

Motor: Participants were presented with visual cues asking them to tap their left or right fingers, squeeze their left or right toes, or move their tongue.

Language: Participants either heard a brief fable (story trials) or an arithmetic problem (math trials) and were subsequently given a two-alternative question about the story or arithmetic problem.

Social: Participants were presented with short video clips of objects that either interacted in some way or moved randomly. Subsequently, participants were asked to decide whether the objects interacted with one another, did not have an interaction, or if they are not sure.

Relational: Participants were presented with different shapes, filled with different textures. In relational trials, participants saw a pair of objects at the top of the screen and a pair at the bottom. They were then asked to decide whether the bottom pair differs along the same dimension (shape or texture) as the top pair. In match trials, participants saw one object at the top and bottom and were asked to decide whether the objects matched on a given dimension.

Emotion: Participants were asked to decide which of two faces presented on the bottom of the screen matches the face at the top of the screen. The faces had an either angry or fearful expression.

fMRI data for each task were provided in a preprocessed format by the Human Connectome Project (HCP S1200 release), WU Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Whole-brain EPI acquisitions were acquired with a 32 channel head coil on a modified 3T Siemens Skyra with TR = 720 ms, TE = 33.1 ms, flip angle = 52 deg, BW = 2,290 Hz/Px, in-plane FOV = 208 x 180 mm, 72 slices, 2.0 mm isotropic voxels with a multi-band acceleration factor of 8. Two runs were acquired, one with a right-to-left and the other with a left-to-right phase encoding (for further methodological details on fMRI data acquisition, see Uğurbil et al. (2013)).

The Human Connectome Project preprocessing pipeline for functional MRI data ("fMRIVolume"; Glasser et al., 2013)) includes the following steps: gradient unwarping, motion correction, fieldmap-based EPI distortion correction, brain-boundary based registration of EPI to structural T1-weighted scan, non-linear registration to MNI152 space, and grand-mean intensity normalization (for further details, see Glasser et al., 2013, Uğurbil et al., 2013).

In addition to the minimal preprocessing of the fMRI data that was performed by the Human Connectome Project, we applied the following preprocessing steps to the fMRI data for all decoding analyses: volume-based smoothing of the fMRI sequences with a 3 mm FWHM Gaussian kernel, linear detrending and standardization of the single voxel signal time-series (resulting in a zero-centered voxel time-series with unit variance) and temporal filtering of the single voxel time-series with a butterworth highpass filter and a cutoff of 128 s, as implemented in Nilearn 0.4.1 (Abraham et al., 2014).

Multi-task dataset

This dataset was first published by Nakai and Nishimoto (2020) and contains the data of six healthy human participants (22-33 yrs, two female, normal vision and hearing,

all right-handed) who repeatedly performed 103 simple naturalistic tasks in the fMRI. These tasks were selected such that they included a variety of different cognitive domains and can be performed without previous experimental training (e.g., participants were asked whether a piece of music is Jazz or whether a penguin is shown on a a presented image; for an overview of all tasks and instructions, see Nakai and Nishimoto, 2020). The experiment consisted of 18 fMRI runs of which 12 were designated as training runs and six as test runs. In each run, 77-83 trials were presented with a duration of 6-12 s per trial. Additionally, a two second feedback (correct, incorrect) for the preceding task was presented 9-13 times per run. Each task had 12 different instances of which eight were used in the training runs and four in the test runs. Importantly, there was no overlap between the training and test instances of each task. The task order was pseudorandomized during the training runs, as some tasks depended on one another. In the six test runs, all tasks were presented in the exact same order. Subjects did not receive any explanation of the tasks prior to the experiment and only underwent a small training on how to use the buttons in the fMRI to indicate their responses (with one response pad with two buttons for each hand). The instruction text of each task was presented with the respective stimuli as a single image during the experiment. All stimuli were shown on a projector screen (21.0 x 15.8 ° of visual angle at 30 Hz). The experiment was performed over three days, with six runs on each day.

The unprocessed fMRI data for this experiment were obtained from the original authors (Nakai and Nishimoto, 2020) through OpenNeuro (`https://openneuro.org/`). Whole-brain EPI acquisitions were acquired with a 32 channel head coil on a 3T Siemens TIM Trio with TR = 2000 ms, TE = 30 ms, flip angle = 62 ° , FOV = 192 x 192 mm, resolution = 2 x 2 mm, MB factor = 3. 72 interleaved axial slices were scanned parallel to the anterior and posterior commissure line (that were each 2.0-mm thick without a gap), using a T2*-weighted gradient-echo multiband echo-planar imaging (MB-EPI) sequence. 275 volumes were obtained for each run. In addition, high-resolution T1-weighted images of the whole brain were acquired from all subjects with a magnetization-prepared rapid acquisition gradient echo sequence (MPRAGE, TR = 2530 ms, TE = 3.26 ms, FA = 9 °, FOV = 256 x 256 mm, voxel size = 1 x 1 x 1 mm).

We preprocessed these data using *fMRIPrep* 20.0.5 (Esteban et al. (2019); Esteban et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502). All details on the individual preprocessing steps for this dataset are reported in Appendix B.1.11.

2.3.2 Conventional analysis approaches

The following will provide an overview of the three conventional approaches to the analysis of fMRI data that we are considering in this work.

General linear model

The General Linear Model (GLM; Friston et al., 1994) represents a univariate brain encoding model (Kriegeskorte and Douglas, 2019, Naselaris et al., 2011). Its goal is to identify an association between cognitive state and brain activity, by predicting the

time series signal of a voxel from a set of experiment predictors:

$$Y = X\beta + \epsilon \tag{2.1}$$

Here, Y presents a $T \times N$ dimensional matrix containing the multivariate time series data of N voxels and T time points. X represents the design matrix, which is composed of $T \times P$ data points, where each column represents one of P predictors. Typically, each predictor represents a variable that is manipulated during the experiment (e.g., stimulus presentation times). β represents a $P \times N$ dimensional matrix of regression coefficients. To mimic the blood-oxygen-level dependent (BOLD) measured by the fMRI, each predictor is first convolved with a hemodynamic response function (HRF; Lindquist et al., 2009), before fitting the β-coefficients to the data. After fitting, the resulting brain map of β-coefficients indicates the estimated contribution of each predictor to the time series signal of each of the N voxels. ϵ represents a $T \times N$ dimensional matrix of error terms. Importantly, the GLM analyzes the time series signal of each voxel independently and thereby includes a separate set of regression coefficients for each voxel in the brain.

Searchlight analysis

The searchlight analysis (Kriegeskorte et al., 2006) is a multivariate brain decoding model (Kriegeskorte and Douglas, 2019, Naselaris et al., 2011). Its goal is to identify an association between cognitive states and brain activity, by probing the ability of a classifier to identify the cognitive states from the activity pattern of a small clusters of voxels. To this end, the entire brain is scanned with a sphere of a given radius (the searchlight) and the performance of the classifier in decoding the cognitive states is evaluated at each location, resulting in a brain map of decoding accuracies. These decoding accuracies indicate how much information about the cognitive state is contained in the activity pattern of the underlying cluster of voxels. We used a radius of 5.6 mm and a linear-kernel Support Vector Machine (SVM) classifier (if not reported otherwise).

Given a training dataset of T data points $[y_t, x_t]_{t=1}^{T}$, where x_t represents the activity pattern of a cluster of voxels at time point t and $y_t \in [-1, 1]$ the corresponding binary class label, the SVM (Cortes and Vapnik, 1995) is generally defined as follows:

$$\hat{y} = sign\left[\sum_{t=1}^{T} \alpha_t y_t \gamma(x, x_t) + b\right] \tag{2.2}$$

Here, α_t and b are positive constants, whereas $\gamma(x, x_t)$ represents the kernel function.

Whole-brain least absolute shrinkage logistic regression

The whole-brain least absolute shrinkage logistic regression (or whole-brain lasso; Grosenick et al., 2013, Wager et al., 2013) represents a whole-brain decoding model (Kriegeskorte and Douglas, 2019, Naselaris et al., 2011). It identifies an association between cognitive state and brain activity, by probing the ability of a logistic model to decode the cognitive state from whole-brain activity (with one logistic coefficient β_i per voxel i in the brain). To reduce the risk of overfitting, resulting from the large number of model

coefficients, the whole-brain lasso applies least absolute shrinkage (or L1) regularization to the likelihood function of the logistic model (Tibshirani, 1996). Thereby, forcing the logistic model to perform automatic variable selection during parameter estimation, resulting in sparse coefficient estimates (i.e., by forcing many coefficient estimates to be exactly 0).

In particular, the optimization problem of the whole-brain lasso can be defined as follows (again, N denotes the number of voxels in the brain, T the number of fMRI sampling time points and $[y_t, x_t]_{t=1}^{T}$ the set of class labels and voxel values of each fMRI sampling time point):

$$\min_{\beta}\{-\sum_{t=1}^{T}\Big[y_t \log \sigma(\beta^T x_t) + (1-y_t)\log(1-\sigma(\beta^T x_t))\Big] + \lambda \sum_{n=1}^{N} |\beta_i|\} \tag{2.3}$$

Here, λ represents the strength of the $L1$ regularization term (with larger values indicating stronger regularization), whereas σ represents the logistic model:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2.4}$$

The resulting set of coefficient estimates β, indicates the weighting of the activity of each voxel i in the decoding decisions $\sigma(x)$ of the logistic model. The whole-brain lasso, as well as closely related decoding approaches (e.g., Beer et al., 2019, Gramfort et al., 2013, Grosenick et al., 2013, McIntosh and Lobaugh, 2004, Ryali et al., 2010), have found widespread application in functional neuroimaging research (e.g., Chang et al., 2015, Toiviainen et al., 2014, Wager et al., 2013).

2.3.3 The DeepLight framework

The DeepLight framework is defined by two central components (for an overview, see Fig. 2.1): It first uses a DL model to decode a cognitive state from a whole-brain fMRI data and subsequently relates the decoded cognitive state and brain activity, by interpreting the decoding decision with the LRP (Bach et al., 2015, Montavon et al., 2017) technique. The LRP technique decomposes the decoding decision of the DL model into the contributions of the activity of the single input voxels to the decision. Importantly, the LRP analysis is performed on the level of a single fMRI volumes, enabling an analysis on several levels of data granularity: From the level of the group down to the level of single subjects, trials, and time points.

Two DeepLight architectures

The first step of the DeepLight analysis is to decode a cognitive state from whole-brain fMRI data by the use of a DL model. Importantly, DeepLight is not restricted to any specific DL model architecture. The following will outline two different DL model architectures that are based on research in computer vision (Donahue et al., 2015, Marban et al., 2019, Tran et al., 2015) (here abbreviated as "2D-DeepLight" and "3D-DeepLight"; see Fig. 2.2).

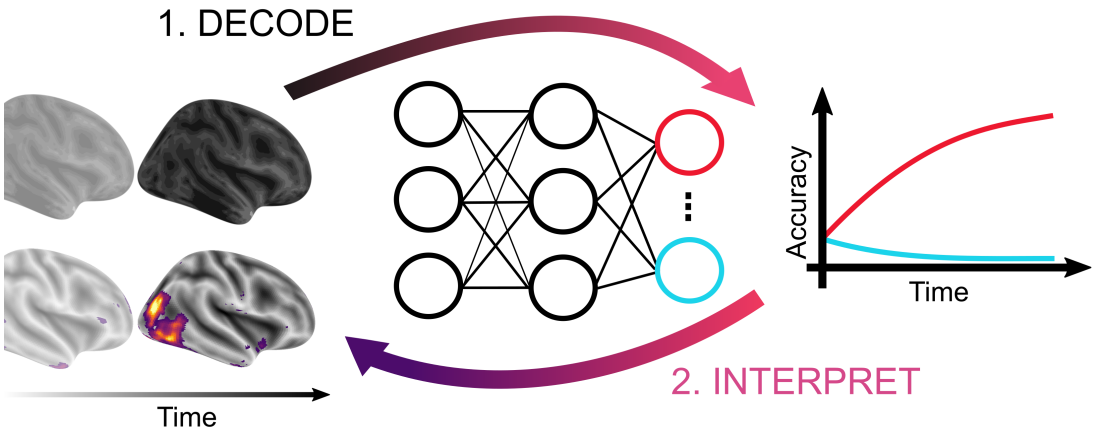

Figure 2.1: The DeepLight framework. First, a DL model is trained to accurately decode (or identify) a set of cognitive states (e.g., viewing the image of a house or a face) from single whole-brain fMRI volumes. Subsequently, DeepLight relates the decoded cognitive state and underlying brain activity by interpreting the decoding decisions of the DL model with layer-wise relevance propagation technique. Here, each decoding decision is decomposed into the contribution (or relevance) of the activity of each voxel to the decision, resulting in a map of relevance values for each voxel of the input fMRI volume.

2D-DeepLight 2D-DeepLight is based on the model used in Thomas et al. (2019a,c) which is composed of three distinct computational modules, namely, a 2D-convolutional feature extractor, an LSTM, and an output unit (for an overview, see Fig. 2.2). First, 2D-DeepLight separates each fMRI volume into a sequence of 2D axial brain slices. These slices are then processed by a 2D-convolutional feature extractor (LeCun et al., 1998), resulting in a sequence of higher-level, and lower-dimensional, slice representations. These higher-level slice representations are fed to an LSTM (Hochreiter and Schmidhuber, 1997), integrating the spatial dependencies of the observed brain activity within and across the axial slices. Lastly, the output unit makes a decoding decision, by projecting the output of the LSTM into a lower-dimensional space, which spans the cognitive states in the data. Here, a probability for each cognitive state is estimated, indicating whether the fMRI volume belongs to each of these states. This combination of convolutional and recurrent DL elements is inspired by previous research which demonstrated that it is well-suited to learn the spatial dependency structure of long sequences of input data (e.g., Donahue et al., 2015, Marban et al., 2019).

The 2D-convolutional feature extractor is composed of a sequence of 2D-convolution layers (LeCun et al., 1998). A 2D-convolution layer consists of a set of kernels (or filters) w that each learns local features of an input image x. These local features are then convolved over the input, resulting in an activation map h, indicating whether a feature is present at each given location of the input:

$$h_{i,j} = g(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n}x_{i+m-1,j+n-1} + b) \tag{2.5}$$

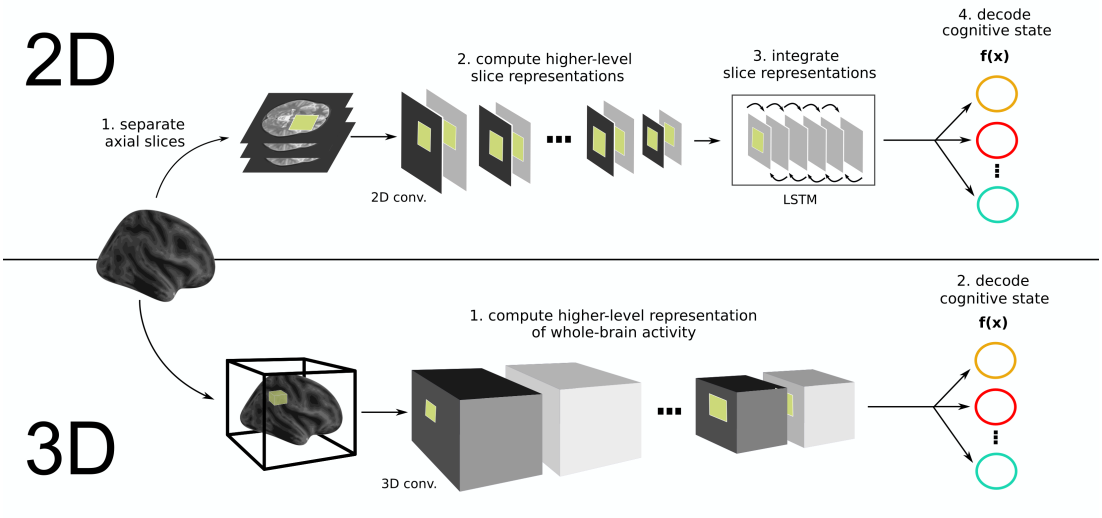

Figure 2.2: Two DeepLight architectures. *2D-DeepLight*: A whole-brain fMRI volume is sliced into a sequence of axial images. These images are passed to a DL model, consisting of a 2D-convolutional feature extractor as well as an LSTM and output unit. First, the 2D-convolutional feature extractor reduces the dimensionality of the axial brain slices through a sequence of 2D-convolution layers. The resulting sequence of higher-level slice representations is fed to a bi-directional LSTM, modeling the spatial dependencies of brain activity within and across brain slices. Lastly, 2D-DeepLight outputs a decoding decision about the cognitive state underlying the fMRI volume, through a softmax output layer with one output neuron per cognitive state in the data. *3D-DeepLight*: A whole-brain fMRI volume is passed to a 3D-convolutional feature extractor, consisting of a sequence of multiple 3D-convolution layers. The 3D-convolutional feature extractor projects the fMRI data into a higher-level, but lower dimensional, representation of whole-brain activity. To make a decoding decision, 3D-DeepLight utilizes an output unit, which is composed of a 1D-convolution and global average pooling layer as well as a softmax function. The 1D-convolution layer maps the higher-level representation of whole-brain activity, resulting from the 3D-convolutional feature extractor, to one representation for each cognitive state in the data. The global average pooling layer and softmax function reduce these to a decoding decision.

Here, b represents the bias of the kernel, while g represents the activation function (see eq. 2.6). The indices m and n represent the row and column index of the kernel matrix, whereas i and j represent the coronal (i.e., row) and saggital (i.e., column) dimensions of the activation map.

Generally, lower-level convolution kernels (closer to the input data) have small receptive fields and are only sensitive to local features of small patches of the input data (e.g., contrasts and orientations). Higher-level convolution kernels, on the other hand, act upon a higher-level representation of the input data, which has already been transformed by a sequence of preceding lower-level convolution kernels. Higher-level kernels thereby integrate the information provided by lower-level convolution kernels, allowing them to identify larger and more complex patterns in the data.

In general, the number of convolution layers and kernels can be adapted for different dimensions and sample sizes of the input fMRI data (see for example, Thomas et al.,

2019a,c). In this work, we utilized two slightly different configurations of 2D-DeepLight, because the datasets of study I and study II differed in both sample size and dimensionality. Note that we applied a brain mask to the fMRI data of study I (for details on the brain mask specification, see Appendix B.1.4) such that it had a dimensionality of $74 \times 92 \times 81$ voxels ($X \times Y \times Z$) instead of the full $91 \times 109 \times 91$ voxels ($X \times Y \times Z$) used in study II. The 2D-DeepLight variant that we used in study I (see section 2.4.1) used eight 2D-convolution layers in the feature extractor, whereas the variant that we used in study II (see section 2.4.2) used 12 2D-convolution layers. For a detailed overview of the two 2D-DeepLight variants, see Appendix B.1.2 and B.1.3.

All convolution kernels in 2D-DeepLight are activated through a rectified linear unit function:

$$g(x) = \max(0, x) \tag{2.6}$$

We generally recommend to move all convolution kernels of the even-numbered layers over the input fMRI data with a stride size of one voxel and all kernels of odd-numbered layers with a stride size of two voxels. The stride size determines the dimensionality of the outputted slice representation. An increasing stride indicates more distance between the applications of the convolution kernels to the input data, thereby reducing the dimensionality of the output representation at the cost of a decreasing sensitivity to differences in the activity patterns of neighboring voxels. Yet, the activity patterns of neighboring voxels are known to be highly correlated, leading to an overall low risk of information loss through a reasonable increase in stride size. To avoid further loss of dimensionality between the convolution layers, we also recommend zero-padding; that is adding zeros to the borders of the inputs to each convolution layer so that the outputs of the convolution layers have the same dimensionality as their inputs, if a stride of one voxel is applied, and only decrease in size when larger strides are used.

To integrate the information provided by the resulting sequence of slice representations into a higher-level representation of the observed whole-brain activity, 2D-DeepLight applies a bi-directional LSTM (Hochreiter and Schmidhuber, 1997), containing two independent LSTM units. Each of the two LSTM units iterates over the entire sequence of input slices, but in reverse order (one from bottom-to-top and the other from top-to-bottom). An LSTM unit contains a hidden cell state C, storing information over an input sequence of length S with elements $x^{(s)}$ and outputs a vector $h^{(s)}$ for each input at sequence step s. The unit has the ability to add and remove information from C through a series of gates. In a first step, the LSTM unit decides which information from the cell state $C^{(s-1)}$ is removed. This is done by a fully-connected logistic layer, the forget gate f:

$$f^{(s)} = \sigma(W_f \cdot x^{(s)} + U_f \cdot h^{(s-1)} + b_f) \tag{2.7}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.8}$$

Here, σ indicates the logistic function, $[W, U]$ the weight matrices of the forget gate, and b the gate's bias. The forget gate outputs a number between 0 and 1 for each

entry in the cell state C. Next, the LSTM unit decides which information is stored in the cell state. This operation contains two elements: the input gate i, which decides which values of $C^{(s-1)}$ will be updated, and a *tanh* layer, which creates a new vector of candidate values $C'^{(s)}$:

$$i^{(s)} = \sigma(W_i \cdot x^{(s)} + U_i \cdot h^{(s-1)} + b_i) \tag{2.9}$$

$$C'^{(s)} = tanh(W_c \cdot x^{(s)} + U_c \cdot h^{(s-1)} + b_c) \tag{2.10}$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.11}$$

Subsequently, the old cell state $C^{(s-1)}$ is updated into the new cell state $C^{(s)}$:

$$C^{(s)} = f^{(s)} C^{(s-1)} + i^{(s)} C'^{(s)} \tag{2.12}$$

Lastly, the LSTM computes its output $h^{(s)}$. Here, the output gate o, decides what part of $C^{(s)}$ will be outputted. Subsequently, $C^{(s)}$ is multiplied by another *tanh* layer to make sure that $h^{(s)}$ is scaled between -1 and 1:

$$o^{(s)} = \sigma(W_o \cdot x^{(s)} + U_o \cdot h^{(s-1)} + b_o) \tag{2.13}$$

$$h^{(s)} = o^{(s)}\ tanh(C^{(s)}) \tag{2.14}$$

To make a decoding decision, both LSTM units forward their output for the last sequence element $h^{(S)}$ to a fully-connected softmax output layer, with one neuron k for each of the K cognitive states in the data:

$$p_k = \frac{e^{h_k^{(S)}}}{\sum_{i=1}^{K} e^{h_i^{(S)}}} \tag{2.15}$$

3D-DeepLight The 3D-DeepLight architecture replaces the combination of 2D-convolution and LSTM that is used by 2D-DeepLight (see Fig. 2.2) with a 3D-convolutional feature extractor, which directly accounts for the three spatial dimensions of whole-brain fMRI data.

A 3D-convolution layer consists of a set of 3D-kernels w that each learn specific features of an input volume x. In contrast to the features learned by 2D-convolution kernels, these features can be three-dimensional (or volumetric). Similar to 2D-convolutions, these features are convolved over the input, resulting in a set of activation maps h, indicating the presence of each of these features at each spatial location of the input volume:

$$h_{i,j,l} = g(\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{z=1}^{Z} w_{m,n,z} x_{i+m-1,j+n-1,l+z-1} + b) \tag{2.16}$$

Again, b represents the bias of the kernel, while g represents the rectified linear unit activation function (see eq. 2.6). The indices m, n, and z index the row, column, and

height of the 3D-convolution kernel, while i, j, and l indicate the coronal (i.e., row), saggital (i.e., column), and axial (i.e., height) dimension of the activation map h.

In this work, we apply 3D-DeepLight only to the fMRI data of study II. Therefore, we only use one configuration of the following 12 3D-convolution layers: conv3-8, conv3-8, conv3-8, conv3-8, conv3-16, conv3-16, conv3-32, conv3-32, conv3-64, conv3-64, conv3-128, conv3-128 [notation: conv(kernel size) - (number of kernels)]. Similar to 2D-DeepLight, this configuration moves all convolution kernels of the even-numbered layers over the input fMRI volume with a stride size of one voxel and all kernels of odd-numbered layers with a stride size of two voxels. 3D-DeepLight further utilizes zero padding, such that the dimensionality of the activation map h only decreases when a stride of more than one voxel is applied.

To make a decoding decision, 3D-DeepLight passes the representation of the feature extractor to an output unit. The output unit is composed of a 1D-convolution layer (with one kernel for each of the cognitive states in the data) as well as a global average pooling layer and softmax function. The purpose of the 1D-convolution layer is to aggregate the information of the C channels of the activation maps h to one activation map for each of the K cognitive states in the data. Subsequently, a global average pooling layer averages over the values of each of the K activation maps and passes its output to a softmax function to obtain a probability estimate p_k for each cognitive state k in the data:

$$h_{i,j,l,k} = g(\sum_{c=1}^{C} w_{k,c} x_{i,j,l,c}) \tag{2.17}$$

$$a_k = \frac{1}{MNZ} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{z=1}^{Z} h_{m,n,z,k} \tag{2.18}$$

$$p_k = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_i}} \tag{2.19}$$

Here, g again indicates the rectified linear unit function (see. eq. 2.6).

Layer-wise relevance propagation

To relate the decoded cognitive state and brain activity, DeepLight utilizes the layer-wise relevance propagation method (LRP; Bach et al., 2015, Montavon et al., 2017, Samek et al., 2020). The goal of the LRP method is to identify the contribution of a single dimension d of an input x (with dimensionality D) to the prediction $f(x)$ that is made by a linear or non-linear classifier f. In the following, the contribution of a single dimension d to the decoding decision is denoted by its relevance R_d. The prediction $f(x)$ of a classifier f can then be decomposed into the sum of the relevance values of each dimension R_d of the input:

$$f(x) = \sum_{d=1}^{D} R_d \tag{2.20}$$

Importantly, LRP assumes that $f(x) > 0$ indicates evidence for the presence of a target and $f(x) < 0$ evidence against the presence of the target. To this end, any $R_d > 0$ can qualitatively be interpreted as evidence that the classifier sees in favor of the presence of the target, while $R_d < 0$ denotes evidence against the presence of the target.

While the relevance of the output neuron at the last layer L is defined as $R_d^{(L)} = f(x)$, the dimension-wise relevance scores on the input neurons are given by $R_d^{(1)}$. The relevance of any other neuron j in any other layer l can then be decomposed into the relevance contributions $R_{i \leftarrow j}^{(l-1,l)}$ of those neurons i in layer $(l-1)$ which provide the inputs to neuron j in layer (l):

$$R_j^{(l)} \approx \sum_{i \in (l-1)} R_{i \leftarrow j}^{(l-1,l)} \tag{2.21}$$

To satisfy eq. 2.20, any definition of the relevance contributions $R_{i \leftarrow j}^{(l-1,l)}$ needs to satisfy the following relevance conservation property between layers (l) and $(l-1)$:

$$\sum_{j \in (l)} R_j^{(l)} \approx \sum_{i \in (l-1)} R_i^{(l-1)} \tag{2.22}$$

For an overview of the different rules that have been proposed to define the relevance contributions $R_{i \leftarrow j}^{(l-1,l)}$, see Bach et al. (2015) and Kohlbrenner et al. (2019). Importantly, DeepLight is not restricted to any specific set of these relevance attribution rules, as long as they follow the principle of relevance conservation defined in eq. 2.22.

Note that in a linear network $f(x) = \sum_i x_{ij}$, in which $R_j = f(x)$, the relevance contributions $R_{i \leftarrow j}$ are directly given by $R_{ij} = x_{ij}$. However, in most DL models, the neuron activation z_j follows a non-linear function of x_j. Importantly, for two of the prominent activation functions, namely, the rectified linear unit function and hyperbolic tangent, the pre-activations x_{ij} provide a sensible way of measuring the contribution of each neuron x_i to R_j (for a more detailed discussion on this issue, see Bach et al., 2015).

LRP for 2D-DeepLight In the context of this work, and in line with the recommendations by Arras et al. (2017a,b), the contributions $R_{i \leftarrow j}^{(l-1,l)}$ for all weighted connections of 2D-DeepLight (see, for example, eq. 2.7, 2.9, 2.13) are defined as:

$$R_{i \leftarrow j}^{(l-1,l)} = \frac{z_{ij}^{(l-1)}}{z_j^{(l-1)} + \epsilon \; sign(z_j^{(l-1)})} R_j^{(l)} \tag{2.23}$$

Here, $z_{ij}^{(l-1)} = x_i^{(l-1)} w_{ij}^{(l-1)}$ (with w indicating the weights and x the input of the weighted connection of layer $(l-1)$) and $z_j^{(l-1)} = \sum_i z_{ij}^{(l-1)} + b_j^{(l-1)}$ (with $b_j^{(l-1)}$ indicating the bias of neuron j), while ϵ represents a stabilizer term that is necessary to avoid numerical degenerations when $z_j^{(l-1)}$ is close to 0 (we set $\epsilon = 0.001$).

Importantly, the LSTM unit of 2D-DeepLight also applies another, multiplicative type of connection (see eq. 2.12 and 2.13). Let $z_j^{(l)}$ be an upper-layer neuron whose value in the forward pass is computed by multiplying two lower-layer neuron values $z_g^{(l-1)}$ and $z_s^{(l-1)}$ such that $z_j^{(l)} = z_g^{(l-1)} z_s^{(l-1)}$. These multiplicative connections occur when one multiplies the outputs of a gate neuron, whose values range between 0 and 1, with an instance of the hidden cell state, which is here referred to as source neuron. For these types of connections, we set the relevances of the gate neuron $R_g^{(l-1)} = 0$ and the relevances of the source neuron $R_s^{(l-1)} = R_j^{(l)}$, where $R_j^{(l)}$ denotes the relevances of the upper layer neuron $z_j^{(l)}$ (as proposed in Arras et al., 2017b). The reasoning behind this rule is that the gate neuron already decides in the forward pass how much of the information contained in the source neuron should be retained to make the classification. Even if this seems to ignore the values of the neurons $z_g^{(l-1)}$ and $z_s^{(l-1)}$ for the redistribution of relevance, these are actually taken into account when computing the value $R_j^{(l)}$ from the relevances of the neurons to which $z_j^{(l)}$ is connected by the weighted connections.

LRP for 3D-DeepLight 3D-DeepLight represents a fully-convolutional neural network, in which the convolution kernels are activated through rectified linear unit functions (ReLU) (see eq. 2.6). Recent empirical work in computer vision (Kohlbrenner et al., 2019) has shown that class discriminability and object localization of the LRP technique can be increased for convolutional network models, by defining the relevance contributions $R_{i \leftarrow j}^{(l-1,l)}$ of all weighted connections of neuron i in layer $(l-1)$ to neuron j in layer (l) as follows:

$$R_{i \leftarrow j}^{(l-1,l)} = (\alpha \frac{z_{ij}^{(l-1)(+)}}{z_j^{(l-1)(+)}} + \beta \frac{z_{ij}^{(l-1)(-)}}{z_j^{(l-1)(-)}}) R_j^{(l)} \tag{2.24}$$

Here, $z_j^{(l-1)(+)} = \sum_i z_{ij}^{(l-1)(+)} + b_j^{(l-1)(+)}$ and $z_j^{(l-1)(-)} = \sum_i z_{ij}^{(l-1)(-)} + b_j^{(l-1)(-)}$, where "+" and "−" indicate the respective positive and negative parts of $z_{ij}^{(l-1)}$ and $b_j^{(l-1)}$. α and β represent two weighting parameters, which allow to scale the contribution of $z_j^{(l-1)(+)}$ and $z_j^{(l-1)(-)}$ to $R_{i \leftarrow j}^{(l-1,l)}$. To satisfy the local conservation property (see eq. 2.22) α and β are restricted to $\alpha + \beta = 1$. In line with the recommendations by Kohlbrenner et al. (2019), we set $\alpha = 2$.

Note that the average pooling layer of the output unit (see eq. 2.17) is a special case of a linearly weighted connection and is thereby subject to the relevance attribution rule defined in eq. 2.23.

2.4 Empirical results

2.4.1 Study I: Comparison to conventional analysis approaches

Decoding performance

In a first study, we compared the performance of DeepLight in decoding a set of cognitive states from fMRI data to the decoding performance of the searchlight analysis and whole-brain lasso (for an overview of these approaches, see section 2.3.2). In this study, we utilized the 2D-DeepLight architecture (see section 2.3.3) and the fMRI data of 100 subjects in the HCP working memory task (for an overview of the working memory task, see section 2.3.1, Appendix B.1.1 and Barch et al., 2013)). In this task, subjects viewed images of either body parts, faces, places or tools in two separate fMRI experiment runs. We first split the fMRI data of these 100 subjects into two distinct training and test datasets (each containing the data of 70 and 30 randomly assigned subjects). All analyses presented throughout the following solely include the data of the 30 subjects contained in the held-out test dataset (if not stated otherwise). The goal of the decoding analysis was to identify the four image categories (i.e., cognitive states) from single fMRI volumes (i.e., TRs). We did not exclude any TR of an experiment block of the four stimulus classes from the decoding analyses. However, we removed all fixation blocks from the decoding analyses (see section 2.3.1 and Appendix B.1.1). Note that the three analysis approaches differ in the number of voxels they include in their analyses. While the searchlight analysis utilizes the data of clusters of multiple voxels (see section 2.3.2) to identify a cognitive state, the whole-brain lasso (see section 2.3.2) and DeepLight (see section 2.3.3) utilize the data of all voxels of an fMRI volume.

After 60 training epochs (see Appendix B.1.6 for an overview of the training procedures and Appendix Fig. B.1 for an overview of the training decoding accuracies), 2D-DeepLight accurately decoded the four cognitive states underlying 68.3% of the fMRI volumes in the held-out test dataset (62.36%, 69.87%, 75.97%, 65.09% for body, face, place and tool, respectively; Fig. 2.3 A). 2D-DeepLight generally performed best at discriminating the body and place (5.1% confusion in the held-out data), face and tool (7.8% confusion in the held-out data), body and tool (9.8% confusion in the held-out data), and face and place (10.4% confusion in the held-out data) stimuli based on the fMRI data, while it did not perform as well in discriminating place and tool, and body and face stimuli (15% confusion in the held-out data, respectively).

Note that 2D-DeepLight's performance in decoding the four cognitive states from the fMRI data varied over the course of an experiment block (Fig. 2.3 D). 2D-DeepLight performed best in the middle and later stages of the experiment block, where the decoding accuracy reaches 80%. This finding is generally in line with the temporal evolution of the haemodynamic response function (HRF; Lindquist et al., 2009) measured by the fMRI (the HRF is known to be strongest 5-10 s after to the onset of the underlying neuronal activity). To further evaluate 2D-DeepLight's performance in decoding the cognitive states from the fMRI data, we compared its performance in decoding these states to the searchlight analysis and whole-brain lasso. For simplicity, we sub-divided this comparison into a separate analysis on the group- and subject-level.

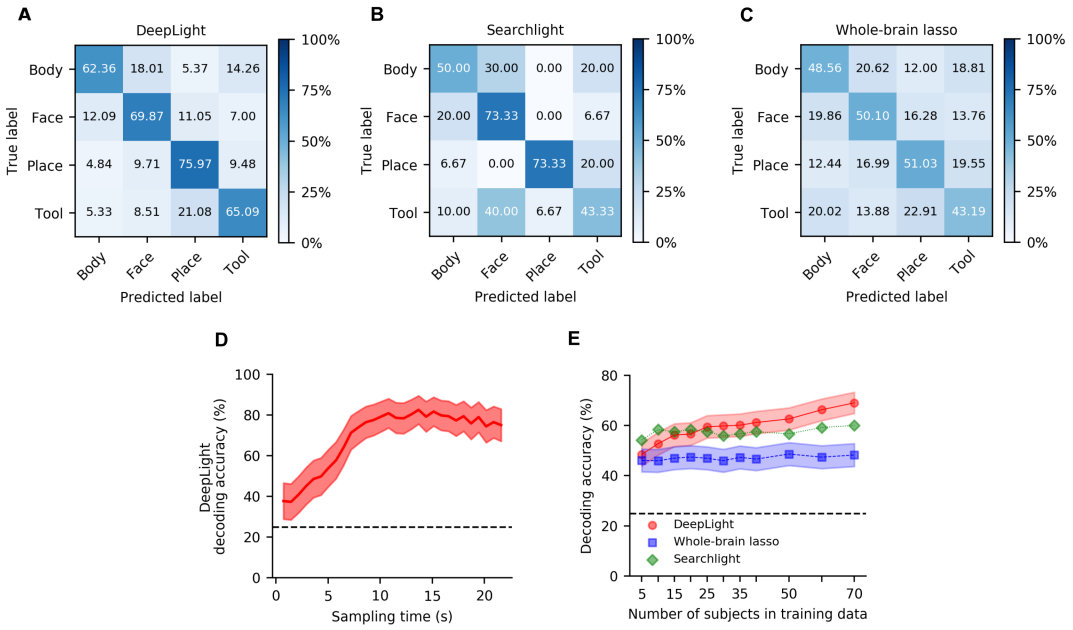

Figure 2.3: Group-level decoding performance comparison of 2D-DeepLight, the searchlight analysis, and whole-brain lasso in the test data of the HCP working memory task (N = 30). A-C: 2D-DeepLight (A) is generally more accurate in identifying the four cognitive states from the fMRI data (as indicated by lower error rates in the confusion matrix), when compared to the searchlight analysis (B) and whole-brain lasso (C). D: 2D-DeepLight's accuracy in identifying the cognitive state of an fMRI volume varies over the course of an experiment block, such that it reaches its maximum 7-10s after block onset. E: 2D-DeepLight's decoding accuracy increases more with growing training dataset sizes (as indicated by the number of subjects in the training dataset), when compared to the searchlight analysis and whole-brain lasso. Lines indicate average decoding accuracies over subjects, with standard errors surrounding them. Colors in E indicate the three analysis approaches.

Group-level comparison: For the group-level comparison, we trained the searchlight analysis and whole-brain lasso on the data of all 70 subjects contained in the training dataset (for details on the training procedures, see Appendix B.1.5). Subsequently, we evaluated their performance in decoding the cognitive states of the 30 subjects in the full held-out test data. 2D-DeepLight clearly outperformed the other approaches in decoding the cognitive states. While the searchlight analysis achieved a decoding accuracy of 60% (Fig. 2.3 B) and the whole-brain lasso a decoding accuracy of 47.97% (Fig. 2.3 C), 2D-DeepLight improved upon these performances by 8.3% ($t(29) = 5.80, P < 0.0001$) and 20.33% ($t(29) = 13.39, P < 0.0001$), respectively. All three decoding approaches generally performed best at discriminating face and place stimuli from the fMRI data (Fig. 2.3 A-C). Similar to 2D-DeepLight, the searchlight analysis and whole-brain lasso also performed well at discriminating body and place stimuli (3.3% and 12.2% confusion for the searchlight analysis and whole-brain lasso, respectively, Fig. 2.3 B-C), while they had more difficulties discriminating body and face

stimuli (25% and 20.2% confusion for the searchlight analysis and whole-brain lasso, respectively, Fig. 2.3 B-C).

A key premise of DL methods, when compared to more traditional decoding approaches, is that their decoding performance improves better with growing datasets. To test this, we repeatedly trained all three decoding approaches on a subset of the training dataset (including the data of 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, and 70 subjects) and validated their performance at each iteration on the full held-out test data (Fig. 2.3 E). Overall, the decoding performance of 2D-DeepLight increased by 0.27% ($t(10) = 10.9, P < 0.0001$) per additional subject in the training dataset, whereas the performance of the whole-brain lasso increased by 0.03% ($t(10) = 3.02, P = 0.015$), and the performance of the searchlight analysis only marginally by 0.04% ($t(10) = 2.08, P = 0.067$). Nevertheless, the searchlight analysis outperformed 2D-DeepLight in decoding the cognitive states from the fMRI data when only little training data were available (here, 10 or less subjects; $t(29) = -4.39, P < 0.0001$). The decoding advantage of 2D-DeepLight, on the other hand, came to light when the data of 50 or more subjects were available in the training dataset ($t(29) = 3.82, P = 0.0006$). 2D-DeepLight consistently outperformed the whole-brain lasso, when it was trained on the data of at least 10 subjects ($t(29) = 5.32, P = 0.0045$).

Subject-level comparison: For the subject-level comparison, we first trained the searchlight analysis and whole-brain lasso on the fMRI data of the first experiment run of a subject from the held-out test dataset (for an overview of the training procedures, see Appendix B.1.5). We then used the data of the second experiment run of the same subject to evaluate their decoding performance (by predicting the cognitive states underlying each fMRI volume of the second experiment run). Importantly, we also decoded the same fMRI volumes with 2D-DeepLight. Note that 2D-DeepLight, in comparison to the other approaches, did not see any data of the subject during the training, as it was solely trained on the data of the 70 subjects in the training dataset. 2D-DeepLight clearly outperformed the other decoding approaches, by decoding the cognitive states more accurately for 28 out of 30 subjects, when compared to the searchlight analysis (while the searchlight analysis achieved a decoding accuracy of 47.2% across subjects, 2D-DeepLight improved upon this performance by 22.4%, with an average decoding accuracy of 69.3%, $t(29) = 11.28, P < 0.0001$; Fig. 2.4 A), and for 29 out of 30 subjects, when compared to the whole-brain lasso (while the whole-brain lasso achieved an average decoding accuracy of 37% across subjects, 2D-DeepLight improved upon this performance by 32.3%; $t(29) = 15.74, P < 0.0001$; Fig. 2.4 B). To further ascertain that the observed differences in decoding performance between the searchlight analysis and 2D-DeepLight did not result from the linearity contained in the Support Vector Machine (SVM; Cortes and Vapnik, 1995) of the searchlight analysis, we replicated our subject-level searchlight analysis, by the use of a non-linear radial basis function kernel (RBF; Cortes and Vapnik, 1995, Müller et al., 2001, Schölkopf et al., 2002) SVM (see Appendix Fig. B.2). However, the decoding accuracies achieved by the RBF-kernel SVM were not meaningfully different from those of the linear-kernel SVM ($t(29) = -1.75, P = 0.09$). Lastly, we also compared the subject-level decoding performance of the whole-brain lasso to that of a recently proposed extension of this

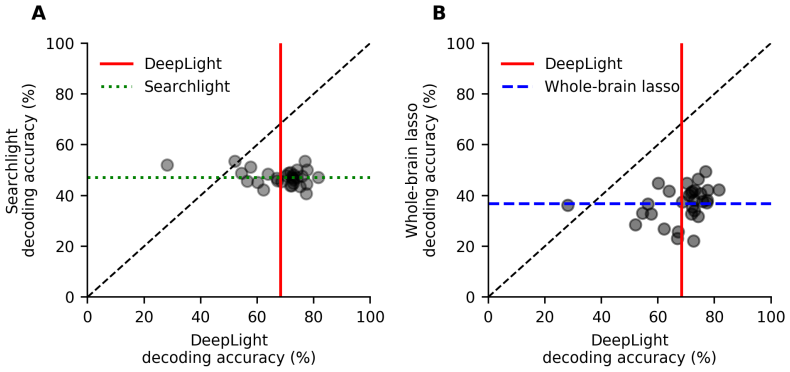

Figure 2.4: Subject-level decoding performance comparison of 2D-DeepLight (red) to the searchlight analysis (A; green) and whole-brain lasso (B; blue) in the test data of the HCP working memory task (N = 30). Black scatter points indicate the average decoding accuracy for a subject. Colored lines indicate the average decoding accuracy across all test subjects.

approach (TV-L1, for methodological details see Gramfort et al., 2013). The TV-L1 approach combines the Least Absolute Shrinkage Regularization (L1; see eq. 2.3) of the whole-brain lasso with an additional Total-Variation (TV) penalty (Michel et al., 2011), to better account for the spatial dependency structure of fMRI data. Yet, we found that the whole-brain lasso performed better at decoding the cognitive states from the subject-level fMRI data than TV-L1 ($t(29) = 3.79, P = 0.0007$; see Appendix Fig. B.3).

Taken together, these findings demonstrate that DeepLight accurately decodes a set of cognitive states from whole-brain fMRI volumes. DeepLight was generally also more accurate in decoding the cognitive states than the searchlight analysis and whole-brain lasso, in a comparison on the level of the individual and the level of the group.

Learned mappings between cognitive states and brain activity

Our previous analyses have shown that 2D-DeepLight has learned a meaningful mapping between the fMRI data and cognitive states of the HCP working memory task, by accurately decoding these states from individual fMRI volumes (see section 2.4.1). Next, we therefore tested 2D-DeepLight's ability to identify the brain areas associated with the cognitive states. To generate a set of subject-level brain maps with 2D-DeepLight, we first decomposed its decoding decisions for each correctly classified fMRI sample of a subject with the LRP method (see section 2.3.3). We further restricted the LRP analysis to those fMRI samples that were collected 5–15 s after the onset of the experiment block, as we expect the HRF (Lindquist et al., 2009) to be strongest within this time period. To then aggregate the resulting set of relevance maps for each decomposed fMRI sample within each cognitive state, we smoothed each relevance map with a 3 mm FWHM Gaussian kernel and averaged all relevance volumes belonging to a cognitive state, resulting in one brain map per subject and cognitive state. Group-level brain maps were then obtained, by averaging these subject-level brain maps for all subjects in the

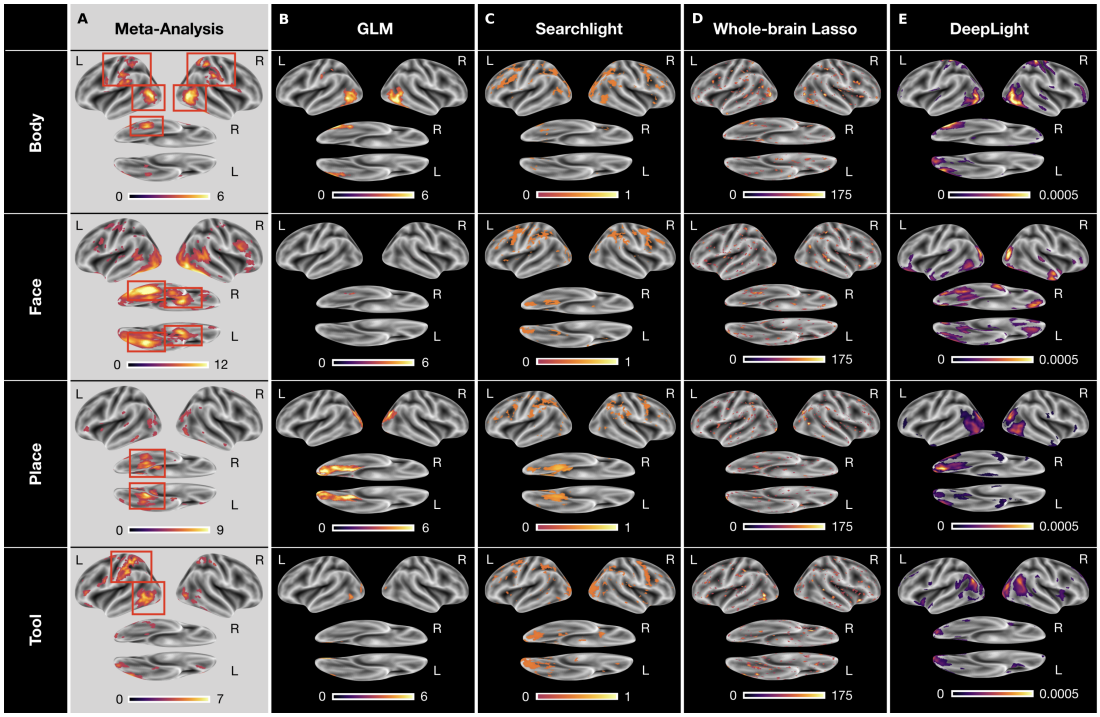

Figure 2.5: Group-level brain maps for each cognitive state in the test data of the HCP working memory task (N = 30). A: Results of a NeuroSynth meta-analysis for the terms "body," "face," "place," and "tools." The brain maps were thresholded at an expected false discovery rate of 0.01. Red boxes highlight the regions-of-interest for each cognitive state. B: Results of the GLM group-level analysis. The brain maps of the GLM analysis were thresholded at an expected false discovery rate of 0.1. C–E: Results of the group-level searchlight analysis (C), whole-brain lasso (D), and 2D-DeepLight (E). The brain maps of the searchlight analysis, whole-brain lasso, and 2D-DeepLight were thresholded at the 90th percentile of their values. Note that the values of the brain maps are on different scales between analysis approaches, due to their different statistical nature. All brain maps are projected onto the inflated cortical surface of the FsAverage5 surface template

held-out test dataset within each cognitive state, resulting in one group-level brain map per cognitive state.

Subsequently, we compared the resulting group- and subject-level brain maps of 2D-DeepLight to those of the GLM, searchlight analysis and whole-brain lasso. Note that due to the diverse statistical nature of the three baseline approaches, the values of their brain maps are on different scales and have different statistical interpretations (for methodological details on the brain maps of the other analysis approaches, see section 2.3.2 and Appendix B.1.5).

To evaluate the quality of the brain maps resulting from each analysis approach, we performed a meta-analysis of the four cognitive states with NeuroSynth (for details on NeuroSynth, see Appendix B.1.9 and Yarkoni et al. (2011)). NeuroSynth provides a database of mappings between cognitive states and brain activity, based on the empirical neuroscience literature. Particularly, the resulting brain maps used here indicate

whether the probability that an article reports a specific brain activation is different, when it includes a specific term (e.g., "face") compared to when it does not. With this meta-analysis, we defined a set of regions-of-interest (ROIs) for each cognitive state (as defined by the terms "body," "face," "place," and "tools"), in which we would expect the various analysis approaches to identify a positive association between the cognitive state and brain activity (for an overview, see Fig. 2.5 A). These ROIs were defined as follows: the upper parts of the middle and inferior temporal gyrus, the postcentral gyrus, as well as the right fusiform gyrus for the body state, the fusiform gyrus (also known as the fusiform face area FFA Haxby et al., 2001, 2000, Heekeren et al., 2004) and amygdala for the face state, the parahippocampal gyrus (or parahippocampal place area PPA; Haxby et al., 2001, 2000, Heekeren et al., 2004) for the place state, and the left upper middle and inferior temporal gyrus as well as the left postcentral gyrus for the tool state. To ensure comparability with the results of the meta-analysis, we restricted all analyses of brain maps to the estimated positive associations between brain activity and cognitive states (i.e., positive relevance values as well as positive GLM and whole-brain lasso coefficients, see section 2.3.2 and Appendix B.1.5). A negative Z-value in the meta-analysis indicates a lower probability that an article reports a specific brain activation when it includes a specific term, compared to when it does not include the term. A negative value in the meta-analysis is therefore conceptually different to negative values in the brain maps of our analyses (e.g., negative relevance values or negative whole-brain lasso coefficients). These can generally be interpreted as evidence against the presence of a cognitive state (or evidence for the presence of any of the other states), given the specific set of cognitive states in our dataset.

Group-level comparison: To determine the voxels that each analysis approach associated with a cognitive state, we defined a threshold for the values of each group-level brain map, indicating those voxels that are associated most strongly with the cognitive state. For the GLM analysis, we thresholded all P-values at an expected false discovery rate (Benjamini and Hochberg, 1995, Genovese et al., 2002) of 0.1 (Fig. 2.5 B). Similarly, for all decoding analyses, we thresholded each brain map at the 90th percentile of its values (Fig. 2.5 C–E). For the whole-brain lasso and DeepLight, the remaining 10 percent of values indicate those brain regions whose activity these approaches generally weight most in their decoding decisions. For the searchlight analysis, the remaining 10 percent indicate those brain regions in which the searchlight analysis achieved the highest decoding accuracy.

All analysis approaches correctly associated activity in the upper parts of the middle and inferior temporal gyrus with body stimuli. The GLM, whole-brain lasso and 2D-DeepLight also correctly associated activity in the right fusiform gyrus with body stimuli. Only 2D-DeepLight correctly associated activity in the postcentral gyrus with body stimuli. The GLM, whole-brain lasso and 2D-DeepLight further all correctly associated activity in the right FFA with face stimuli. None of the approaches, however, associated activity in the left FFA with face stimuli. Interestingly, the searchlight analysis did not associate the FFA with face stimuli at all. All analysis approaches also correctly associated activity in the PPA with place stimuli. Lastly, for tool stimuli, the GLM and whole-brain lasso correctly associated activity in the left inferior temporal

sulcus with stimuli of this class. The searchlight analysis and whole-brain lasso only did so marginally. None of the approaches associated activity in the left postcentral gyrus with tool stimuli.

Overall, 2D-DeepLight's group-level brain maps accurately associated each of the ROIs with their respective cognitive states. Interestingly, 2D-DeepLight also associated a set of additional brain regions with the face and tool stimulus classes that were not identified by the other analysis approaches (see Fig. 2.5 E). For face stimuli, these regions are the orbitofrontal cortex and temporal pole. While the temporal pole has been shown to be involved in the ability of an individual to infer the desires, intentions and beliefs of others (theory-of-mind; for a detailed review, see Olson et al., 2007), the orbitofrontal cortex has been associated with the processing of emotions in the faces of others (for a detailed review, see Adolphs, 2002). For tool stimuli, 2D-DeepLight additionally utilized the activity of the temporoparietal junction (TPJ) to decode these stimuli. The TPJ has been shown to be associated with the ability of an individual to discriminate self-produced actions and the actions produced by others and is generally regarded of as a central hub for the integration of body-related information (for a detailed review, see Decety and Grèzes, 2006). Although it is not clear why only 2D-DeepLight associated these brain regions with the face and tool stimulus classes, their assumed functional roles do not contradict this association.

Subject-level comparison: The goal of the subject-level analysis was to test the ability of each analysis approach to identify the physiologically appropriate associations between brain activity and cognitive state on the level of each individual.

To quantify the similarity between the subject-level brain maps and the results of the meta-analysis, we defined a similarity measure. Given a target brain map (e.g., the results of our meta-analysis), this measure tests for each voxel in the brain whether a source brain map (e.g., the results of our subject-level analyses) correctly associates this voxel's activity with the cognitive state (true positive), falsely associates the voxel's activity with the cognitive state (false positives) or falsely does not associate the voxel's activity with the cognitive state (false negatives). Particularly, we derived this measure from the well-known F1-score in machine learning (see Appendix B.1.10). The benefit of the F1-score, when compared to simply computing the ratio of correctly classified voxels in the brain, is that it specifically considers the brain map's precision and recall and is thereby robust to the overall size of the ROIs in the target brain map. Here, precision describes the fraction of true positives from the total number of voxels that are associated with a cognitive state in the source brain map. Recall, on the other hand, describes the fraction of true positives from the overall number of voxels that are associated with a cognitive state in the target brain map. Generally, an F1-score of 1 indicates that the brain map has both, perfect precision and recall with respect to the target, whereas the F1-score is worst at 0.

To obtain an F1-score for each subject-level brain map (for details on the estimation of subject-level brain maps with the three baseline analysis approaches, see Appendix B.1.5), we again thresholded each individual brain map. For the GLM, we defined all voxels with $P > 0.005$ (uncorrected) as not associated with the cognitive state and all others as associated with the cognitive state. For the searchlight analysis, whole-brain

lasso, and 2D-DeepLight, we defined all voxels with a value below the 90th percentile of the values within the brain map as not associated with the cognitive state and all others as associated with the cognitive state.

Overall, 2D-DeepLight's subject-level brain maps had meaningfully larger F1-scores for the body, face, and place stimulus classes, when compared to those of the GLM ($t(29) = 10.46, P < 0.0001$ for body stimuli, Appendix Fig. B.4 A; $t(29) = 13.04, P < 0.0001$ for face stimuli, Appendix Fig. B.4 D; $t(29) = 9.26, P < 0.0001$ for place stimuli, Appendix Fig. B.4 G), searchlight analysis ($t(29) = 13.26, P < 0.0001$ for body stimuli, Appendix Fig. B.4 B; $t(29) = 8.57, P < 0.0001$ for face stimuli, Appendix Fig. B.4 E; $t(29) = 4.25, P = 0.0002$, for place stimuli, Appendix Fig. B.4 H), and whole-brain lasso ($t(29) = 20.93, P < 0.0001$ for body stimuli, Appendix Fig. B.4 C; $t(29) = 48.32, P < 0.0001$ for face stimuli, Appendix Fig. B.4 F; $t(29) = 22.43, P < 0.0001$, for place stimuli, Appendix Fig. B.4 I). For tool stimuli, the GLM and searchlight analysis generally achieved higher subject-level F1-scores than DeepLight ($t(29) = -8.19, P < 0.0001$, Appendix Fig. B.4 J; $t(29) = -4.39, P = 0.0001$, Appendix Fig. B.4 K for the GLM and searchlight, respectively), whereas 2D-DeepLight outperformed the whole-brain lasso ($t(29) = 18.31, P < 0.0001$, Appendix Fig. B.4 L).

To ascertain that the results of this comparison were not dependent on the thresholds that we chose, we replicated the comparison for each combination of the 85th, 90th, and 95th percentile threshold for the brain maps of the searchlight analysis, whole-brain lasso, and 2D-DeepLight, as well as a P-threshold of 0.05, 0.005, 0.0005, and 0.00005 for the brain maps of the GLM. Within all combinations of percentile values and P-thresholds, the presented results of the F1-comparison were generally stable (see Appendix Table B.3, B.4, B.5, B.6).

In summary, these findings demonstrate that 2D-DeepLight identifies a biologically plausible association between fMRI data and cognitive states by first learning to accurately decode these states and subsequently interpreting its decoding decisions with the LRP technique. 2D-DeepLight's learned mapping between fMRI data and cognitive states of the HCP working memory task generally matches the results of a meta-analysis; on the level of the individual subject as well as on the level of the group.

Capturing the temporo-spatial variability of brain activity

To probe 2D-DeepLight's ability to analyze single time points, we also studied the distribution of relevance values over the course of a single experiment block of the HCP working memory task (see Fig. 2.6). In particular, we plotted this distribution as a function of the fMRI sampling-time over all subjects for the first experiment block of the face and place stimulus classes in the second experiment run (for details on the experiment paradigm, see section 2.3.1). We restricted this analysis to the face and place stimulus classes, as the neural networks involved in processing face and place stimuli, respectively, have been widely characterized (e.g., Haxby et al., 2001, 2000, Heekeren et al., 2004).

In the beginning of the experiment block, 2D-DeepLight was generally uncertain which cognitive state the observed brain samples belonged to, as it assigned similar probabilities to each of the cognitive states considered (Fig. 2.6 A, B). As time pro-

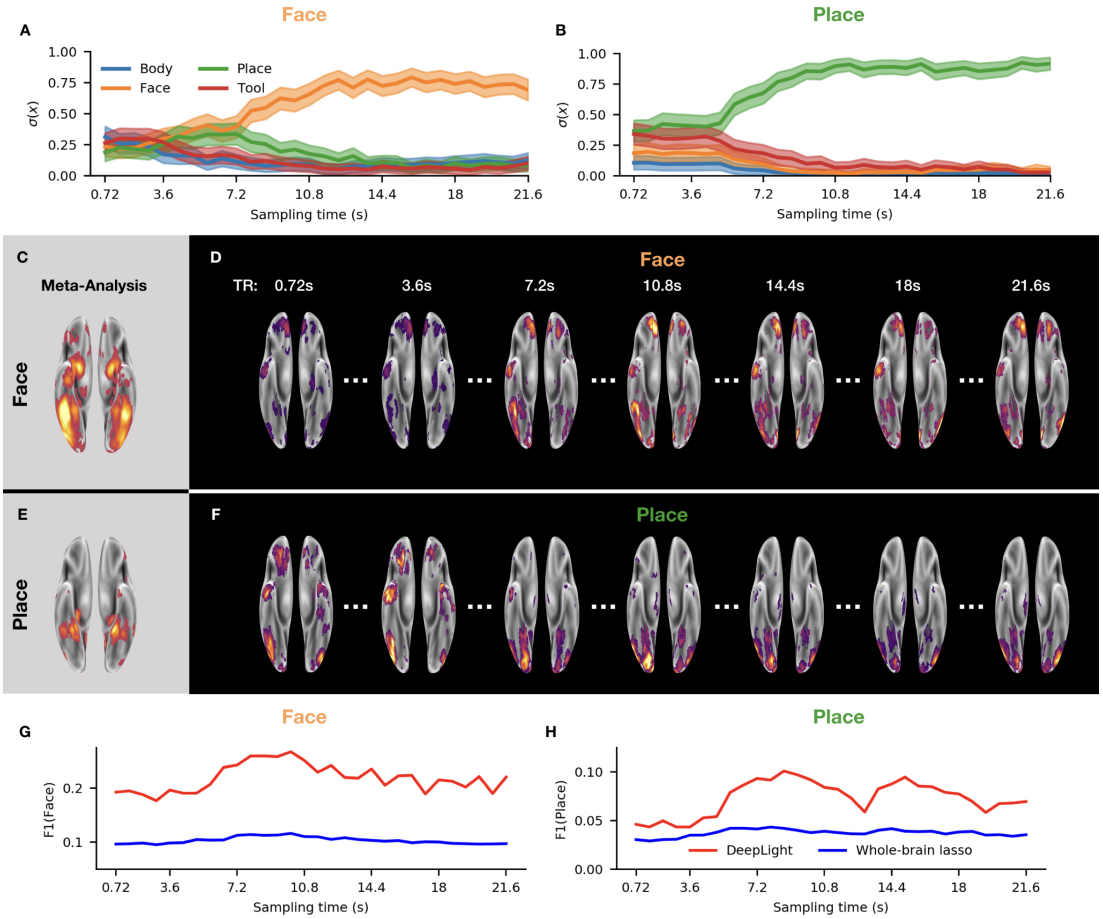

Figure 2.6: DeepLight analysis of the temporo-spatial distribution of brain activity in the first experiment block of the face and place stimulus classes in the test dataset of the HCP working memory task (N = 30). A, B: As time increases, 2D-DeepLight is more certain that an fMRI volume belongs to the face (A) and place (B) stimulus classes. C, E: Results of a meta-analysis with the NeuroSynth database for the face and place stimulus classes (for details on the meta-analysis, see Appendix B.1.9). D, F: As time increases, 2D-DeepLight assigns more relevance to the ROIs of both stimulus classes. Each group-level brain map is computed as an average over the relevance maps of each test subject for this time point and thresholded at the 90th percentile of its values. All brain maps are projected onto the inflated cortical surface of the FsAverage surface template (Fischl, 2012). G, H: F1-score for each time point. The F1-score quantifies the similarity between a brain map and the results of the meta-analysis (C,E) (with larger values indicating more similarity; for further details on the F1-score, see section 2.4 and Appendix B.1.10). Red indicates the results of the F1-score comparison for the brain maps of 2D-DeepLight, whereas blue indicates the results of this comparison for the brain maps of the whole-brain lasso analysis. See the main text for further details on this analysis.

gressed, however, 2D-DeepLight's certainty increased and it correctly identified the cognitive state underlying the fMRI samples. At the same time, it started assigning more relevance to the target ROIs of the face and place stimulus classes (Fig. 2.6 C–F), as

indicated by the increasing F1-scores resulting from a comparison of the brain maps at each sampling time point with the results of the meta-analysis (Fig. 2.6 G, H; all brain maps were again thresholded at the 90th percentile for this comparison). Interestingly, the relevances started peaking in the target ROIs 5 s after the onset of the experiment block. The temporal evolution of the relevances thereby mimics the haemodynamic response measured by the fMRI (Lindquist et al., 2009).

To further evaluate the results of this analysis, we replicated it by the use of the whole-brain lasso group-level decoding model (see section 2.3.2 and Appendix B.1.5). In particular, we multiplied the fMRI samples of all test subjects collected at each sampling time point with the coefficient estimates of the whole-brain lasso group-level model. Subsequently, we averaged the resulting weighted fMRI samples within each sampling time point depicted in Fig. 2.6 G-H and computed an F1-score for a comparison of the resulting average brain maps with the results of the meta-analysis (as described in section 2.4 and Appendix B.1.10). Interestingly, we found that the F1-scores of the whole-brain lasso analysis varied much less over the sequence of fMRI samples and were throughout lower than those of 2D-DeepLight. Thereby, indicating that the brain maps of the whole-brain lasso analysis exhibit comparably little variability over the course of an experiment block with respect to the target ROIs defined for the face and place stimulus classes.

Taken together, these findings demonstrate that 2D-DeepLight can capture the temporo-spatial variability of brain activity, as the distribution of its relevance values resembles the haemodynamic response. Importantly, 2D-DeepLight was trained on the level of individual fMRI volumes, it therefore does not know about the temporal distribution of brain activity. These findings thereby demonstrate 2D-DeepLight's ability to associate a cognitive state with variable patterns of brain activity.

2.4.2 Study II: Transfer learning with fMRI data

Pre-training

To explore whether transfer learning is beneficial for the application of DL models to the decoding of cognitive states from whole-brain fMRI data, we trained a variant of each DeepLight architecture (2D and 3D, see section 2.3.3) on a large dataset of the HCP (for details on the model specifications, see Appendix B.1.3; for details on the training procedures, see Appendix B.1.7). Specifically, the pre-training dataset included the fMRI data of 450 participants in six of the seven HCP experiment tasks (all tasks except for the working memory task; for an overview of the individual experiment tasks, see section 2.3.1 and Appendix B.1.1).

To evaluate the ability of the 2D- and 3D-DeepLight architectures to identify the cognitive states from individual fMRI volumes, we further divided the data of each pre-training task into distinct training and validation datasets, by designating the fMRI data of 400 randomly selected subjects as training data and the data of the remaining 50 subjects as a validation dataset.

During pre-training, the output layer of both DeepLight architectures contained 16 neurons, one for each cognitive state of each task in the pre-training dataset (for an overview of the individual cognitive states, see Table 2.1). The DeepLight variants

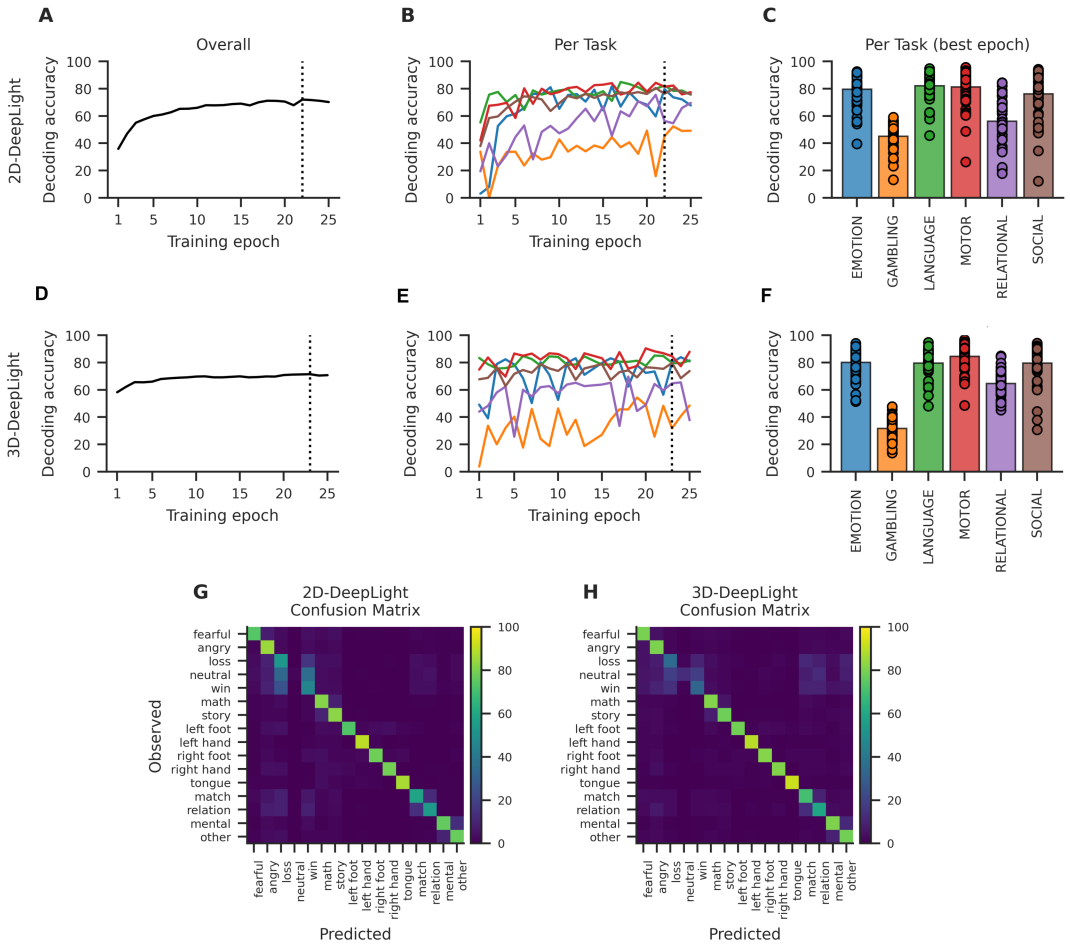

Figure 2.7: Pre-training decoding accuracy of the 2D- and 3D-DeepLight architectures. A-B, D-E: During pre-training, the 2D- (A-B) and 3D-DeepLight (D-E) variants learned to accurately decode the cognitive states from the validation data of the pre-training dataset (with 50 subjects per task). Dashed vertical lines indicate the training epoch with the maximum overall decoding accuracy in the validation data. C, F: Decoding accuracies for each task at the training epoch with the highest overall decoding accuracy in the validation dataset (dashed lines in A and D). Bar heights indicate overall decoding accuracies, with scatter points indicating individual subjects. G, H: Confusion matrices for the 2D- (G) and 3D- (H) DeepLight variants for the training epoch with the highest overall decoding accuracy in the validation dataset. Brighter yellow colors indicate fewer errors.

therefore had no knowledge of the underlying experiment task structure and were solely trained to identify an individual's cognitive state from a single fMRI volume. We trained both DeepLight variants for a period of 25 epochs (Fig. 2.7 A, D). We defined an epoch as an iteration over the entire training data of the pre-training dataset.

Both DeepLight architectures performed well in decoding the cognitive states from the fMRI data of the pre-training dataset (Fig. 2.7 A, D): The 2D-DeepLight variant achieved its highest decoding accuracy in the validation data after 22 training epochs

(72.01%; Fig. 2.7 A), whereas the 3D-DeepLight variant achieved its highest decoding accuracy after 23 training epochs (71.47%; Fig. 2.7 D). Note that these performances were statistically not meaningfully different from one another ($t(299) = -0.17, P = 0.86$).

Both DeepLight architectures generally performed best at identifying the cognitive states of the motor (2D: 81.43%, 3D: 84.64%; Fig. 2.7 B-C, E-F), language (2D: 82.30%, 3D: 79.74%; Fig. 2.7 B-C, E-F), emotion (2D: 79.66%, 3D: 80.16%; Fig. 2.7 B-C, E-F), and social (2D: 76.25%, 3D: 79.68%; Fig. 2.7 B-C, E-F) experiment tasks (see Fig. 2.7 B-C, E-F). They did not perform as well in the relational experiment task (2D: 56.29%, 3D: 64.75%; Fig. 2.7 B-C, E-F) and struggled to accurately decode the cognitive states of the gambling task (2D: 45.12%, 3D: 31.80%; Fig. 2.7 B-C, E-F).

Interestingly, both architectures generally exhibited little confusion between the cognitive states of different tasks (with the exception of the gambling task; see Fig. 2.7 G-H), indicating that they were able to correctly group the cognitive states of the six tasks without receiving any explicit information about the task structure during training.

In summary, both DeepLight architectures (2D and 3D) learned a meaningful mapping between the fMRI data and cognitive states of five out of the six HCP pre-training tasks (with the exception of the gambling task) by accurately decoding these states from individual whole-brain fMRI volumes.

Transfer to a new experiment task

Our results demonstrate that the two pre-trained DeepLight variants (2D and 3D; see section 2.3.3) in general accurately decode the cognitive states of our pre-training dataset (which includes six out of the seven HCP experiment tasks; see Fig. 2.7 C, G). We were therefore interested to evaluate whether the pre-trained models (which achieved the highest validation decoding accuracies during pre-training; dashed lines in Fig. 2.7 A, E), perform better in identifying the cognitive states of the left-out HCP experiment task (the working memory task; see section 2.3.1) than two respective model variants with the same architecture that are trained from scratch.

In a first step of this analysis, we compared the performance of two fine-tuning approaches (see section 2.1.4), by testing whether a model that freezes the pre-trained weights during fine-tuning performs better than a model variant that continues to train these weights. To do this, we first initialized the weights of two identical variants of each architecture (for an overview of the model architectures, see Appendix B.1.3) to the weights of the pre-trained models (except for weights of the output units, which now included four instead of 16 neurons; see section 2.3.3). We then held the pre-trained weights of one variant constant during fine-tuning, while we allowed the weights of the respective other variant to change. After 50 training epochs (for an overview of the training procedures, see Appendix B.1.7), the model variants whose weights were allowed to change clearly outperformed the model variants with frozen weights (see Appendix Fig. B.5). While the models that were not allowed to change the pre-trained weights achieved a final decoding accuracy of 54.73% (2D) and 64.50% (3D) in the validation data of the HCP working memory task, the model variants that were allowed

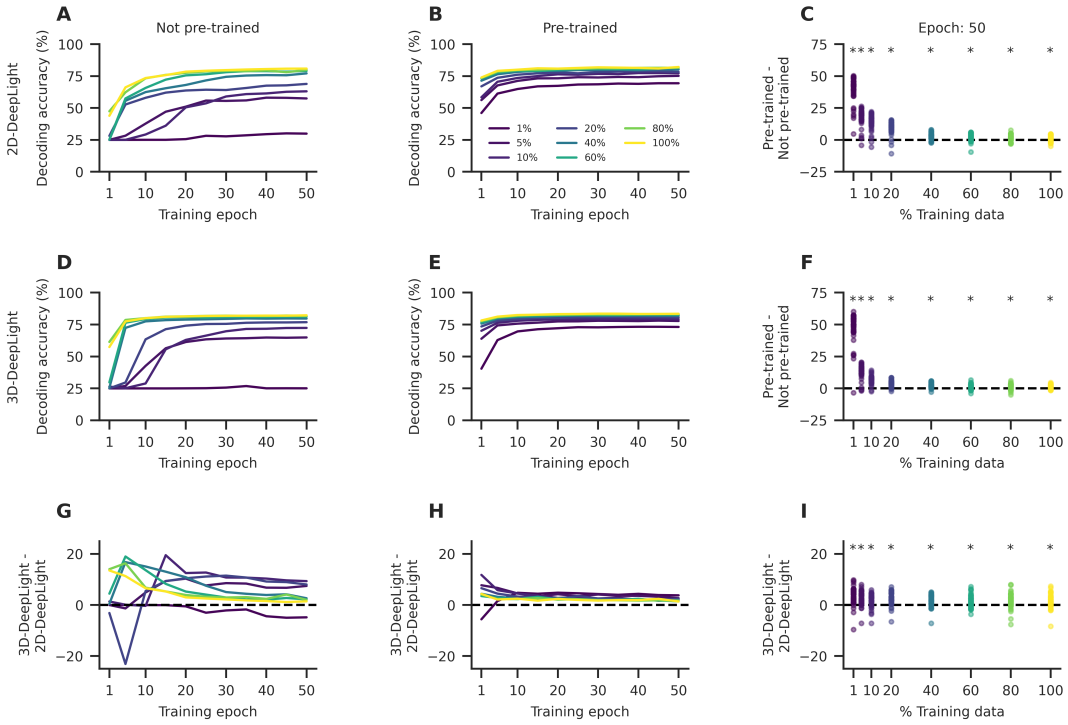

Figure 2.8: Decoding performance of a pre-trained and not pre-trained variant of the 2D- (A-C) and 3D-DeepLight (D-F) architectures in the validation data of the HCP working memory task (N = 50). We repeatedly trained the pre-trained (A, D) and not pre-trained (B, E) DepeLight variants on the fMRI data of 1%, 5%, 10%, 20%, 40%, 60%, 80%, and 100% of the training dataset of the HCP working memory task (N = 400). After 50 training epochs (C, F), the pre-trained DeepLight variants consistently achieved higher decoding accuracies than their not pre-trained counterparts in the validation data. G-I: Difference between the decoding accuracies of the 2D- (A-C) and 3D-DeepLight (D-F) architectures. In general, the 3D-DeepLight architectures achieved higher decoding accuracies than their 2D counterparts. Lines indicate decoding accuracies in the validation data of the working memory task, while scatter points indicate individual subjects. Stars indicate that the distribution of subject decoding accuracies is meaningfully different from 0 in a two-sided t-test (Bonferroni-corrected, such that $P \leqslant 0.05/8$). Colors indicate the different fractions of the training data.

to train these weights achieved a final decoding accuracy of 82.10% (2D) and 83.45% (3D). In all other transfer learning analyses, we therefore allowed the models to change the pre-trained weights during fine-tuning.

In a second step of this analysis, we compared the performance of the pre-trained models to those of two variants that were trained from scratch (with weights initialized according to the random uniform initializiation scheme proposed by Glorot and Bengio, 2010). After 50 training epochs, the model variants that were not pre-trained achieved a final decoding accuracy of 80.83% (2D; Fig. 2.8 A) and 82.23% (3D; Fig. 2.8 D) in the validation data of the HCP working memory task and thereby performed -1.27% ($t(49) = -4.047, P = 0.00018$)) and -1.22% $t(49) = -5.74, P < 0.0001$) worse than

their pre-trained counterparts (Fig. 2.8 C, F).

To also test how the pre-trained and not pre-trained model variants compared when applied to smaller fractions of the full training dataset of the HCP working memory task (N = 400), we trained the pre-trained and not pre-trained variant of each architecture on the data of 1%, 5%, 10%, 20%, 40%, 60%, and 80% of the full training dataset (for an overview of the training procedures, see Appendix B.1.7). We always evaluated the decoding performance of each model on the data of all 50 subjects in the validation dataset.

The pre-trained DeepLight variants consistently achieved higher decoding accuracies than the models that were not pre-trained (Fig. 2.8 C, F). Note that the pre-trained models were able to correctly identify the cognitive state underlying 69.34% (2D) and 73.10% (3D) of the fMRI volumes of the validation dataset, when they were trained on 1% of the training dataset (equal to a dataset of four subjects). The DeepLight variants that were not pre-trained, on the other hand, achieved a decoding accuracy of 29.87% (2D) and 25.02% (3D) when trained on 1% of the training data (thereby performing meaningfully worse than their pre-trained counterparts; the difference in decoding accuracy between the pre-trained and not pre-trained models was $-39,47\%$, $t(49) = -29.13, P < 0.0001$); and $-48,08\%$, $t(49) = -30.20, P < 0.0001$ for the 2D- and 3D architectures respectively). Similarly, the pre-trained DeepLight variants that were fine-tuned on 40% of the training dataset achieved a decoding performance that was as good as the performance of the not pre-trained DeepLight variants that were trained on the data of all 400 subjects in the training dataset (2D: 80.83% (not pre-trained; 100%) - 80.11% (pre-trained; 40%): $t(49) = 2.57, P = 0.013$; 3D: 82.23% (not pre-trained; 100%) - 81.96% (pre-trained; 40%): $t(49) = 1.58, P = 0.12$; Fig. 2.8 A-B, D-E)

In general, the 3D-DeepLight variants were more accurate in identifying the cognitive states from the fMRI data than their 2D counterparts (Fig. 2.8 G - I). While the 3D-DeepLight variants also exhibited faster learning (by achieving higher decoding accuracies earlier in the training), we refrain from interpreting this finding further, as the two DeepLight architectures were trained with slightly different learning rates (for details on the training procedures, see Appendix B.1.7).

Taken together, a DeepLight variant that is pre-trained on the data of six out of the seven HCP experiment tasks generally learns faster and achieves higher decoding accuracies, while also requiring less training data, than a model variant that is trained from scratch, when both are applied to the fMRI data of the left-out seventh HCP experiment task.

The two DeepLight architectures (2D and 3D) have learned a meaningful mapping between the fMRI data and cognitive states of the HCP working memory task, as demonstrated by their ability to accurately decode these states from the fMRI data (Fig. 2.8). Next, we were therefore interested in also comparing their learned mappings between the cognitive states and fMRI data. To this end, we decomposed the decoding decisions of each architecture for each volume in the validation data of the HCP working memory task with the LRP technique (see section 2.3.3). For this comparison, we used the two DeepLight variants that were previously pre-trained (see section 2.4.2), as these generally performed best in identifying the cognitive states of the HCP working

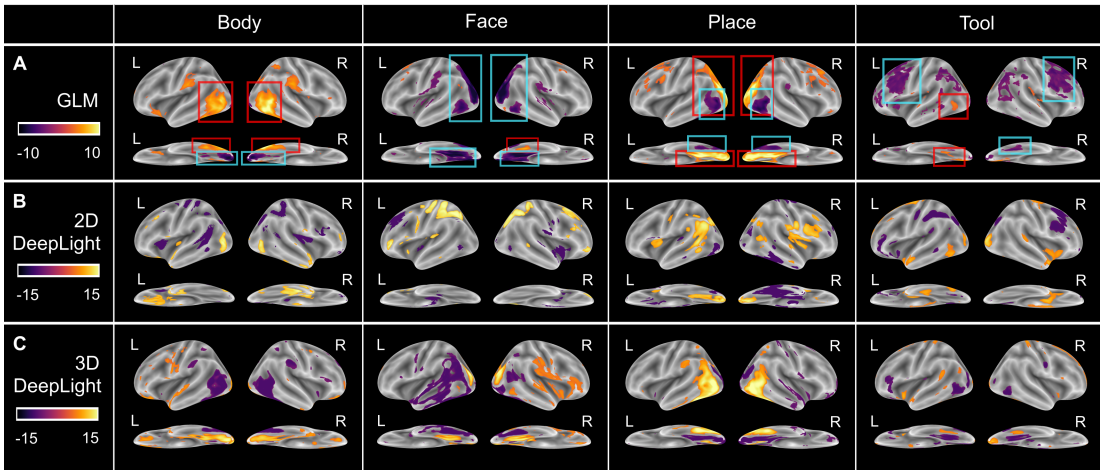

Figure 2.9: Comparison of the learned mappings between the cognitive states and fMRI data of the HCP working memory task (N = 50). A: We first computed a random-effects GLM analysis of the 50 subjects in the validation dataset of the HCP working memory task (by contrasting each cognitive state with all others; see Appendix B.1.8). Based on this analysis, we defined a set of regions-of-interest (ROIs) for each cognitive state, by thresholding the brainmaps of the GLM analysis at an expected false discovery rate of 0.05. Brain regions who exhibited a meaningful positive association with a cognitive state in the GLM analysis are marked in red, whereas brain regions who exhibited a meaningful negative association are marked in blue. B-C: Results of a random-effects GLM analysis of the relevance values of the 2D- (B) and 3D-DeepLight (C) variants for each fMRI sample of the validation dataset (for details on the GLM analysis, see Appendix B.1.8). We thresholded the results of the relevance GLM analysis at the 10th and 90th percentile of the Z-values within each cognitive state, because the relevance data generally have a higher signal-to-noise ratio than the underlying fMRI data (the corresponding false-discovery rates were below 0.01 for all cognitive states). All brain maps are projected onto the inflated cortical surface of the FsAverage template (Fischl, 2012).

memory task (Fig. 2.8). To identify the brain regions whose activity each of the two DeepLight architectures associated most strongly with each of the four cognitive states, we performed a random effects GLM analysis (Friston et al., 2005) of the resulting relevance data (contrasting each cognitive state against all others; see Appendix B.1.8).

As in our previous analysis of the HCP working memory task (see section 2.4), we also performed an additional GLM analysis of the underlying fMRI data (for an overview of the GLM analysis, see Appendix B.1.8). This allowed us to define a set of regions-of-interest (ROIs) whose activity we would expect to be associated with each of the four cognitive states. In contrast to our previous analysis, we did not restrict the ROIs to only positive associations (indicating that the activity of a brain region is more pronounced in one cognitive state than in the others), but also included meaningful negative associations (indicating that the activity of a brain regions is generally weaker in a cognitive state than in the others), as both of these types of associations convey meaningful information about the presence of a cognitive state in a decoding analysis. We defined these associations for each cognitive state by thresholding the P-values of the GLM analysis at a false discovery rate of 0.05 (Fig. 2.9 A). In this analysis, the

upper parts of the middle and inferior temporal gyrus were positively associated with the body state, while the occipital pole exhibited a meaningful negative association with the body state. The right fusiform gyrus (or fusiform face area FFA; Heekeren et al., 2004) exhibited a meaningful positive association with the face state, while the dorsal parts of the occipital cortex, lower parts of the posterior parietal cortex, upper parts of the inferior temporal gyrus, and the parahippocampal gyrus (or parahippocampal place area PPA; Heekeren et al., 2004) exhibited a meaningful negative association with the face state. The PPA, occipital pole, dorsal occipital cortex, and lower parts of the posterior parietal cortex exhibited a meaningful positive association with the place state, whereas the upper parts of the middle and inferior temporal gyrus, and the fusiform gyrus exhibited a meaningful negative association with the place state. Lastly, the upper parts of the left middle and inferior temporal gyrus as well as parts of the left parahippocampal gyrus exhibited a positive association with the tool state, while the right fusiform gyrus and large parts of the dorsolateral prefrontal cortex exhibited a negative association with the tool state.

The learned mapping between cognitive states and fMRI data of 2D-DeepLight was largely in line with the ROIs of our GLM analysis (Fig. 2.9 B). 2D-DeepLight utilized activity of the upper middle and inferior temporal gyrus to identify the body state, activity of the posterior parietal cortex to identify the face state, activity of the occipital poles and parahippocampal gyrus to identify the place state, and activity of the upper parts of the left middle and inferior temporal gyrus to identify the tool state. Interestingly, 2D-DeepLight did not utilize activity of the right FFA to identify the face state, but instead activity of parts of the precentral and postcentral gyrus, in addition to the posterior parietal cortex. Note that the posterior parietal cortex exhibited a meaningful negative association with the face state in our baseline GLM analysis (Fig. 2.9 A). 2D-DeepLight also utilized activity of the left ventromedial prefrontal cortex and temporal poles to identify the body state, activity of the left temporoparietal junction to identify the place state, and activity of parts of the occipital cortex to identify the tool state. The overall set of brain regions that 2D-DeepLight associated with each of the four cognitive states is thereby largely in line with the set of brain regions of our previous analysis of 2D-DeepLight's decoding decisions for the HCP working memory task (see Fig. 2.5 E). Interestingly, the combination of brain regions whose activity 2D-DeepLight associated with each of the four cognitive states was slightly different between both analyses, indicating that 2D-DeepLight is able to identify the four cognitive states through the activity of different combinations from this overall set of brain regions.

Similarly, the overall set of brain regions that 3D-DeepLight associated with each of the four cognitive states was generally in line with our ROIs (Fig. 2.9 A, C). Interestingly, the sign of 3D-DeepLight's brainmaps was inverted, such that 3D-DeepLight learned to identify each cognitive state through the activity of those brain regions that exhibit meaningfully less activity in this state compared to the others in our baseline GLM analysis (Fig. 2.9 A). Specifically, 3D-DeepLight utilized the activity of the parahippocampal gyrus and occipital pole to identify the body state, activity of the parahippocampal gyrus, dorsal occipital cortex, and the posterior part of the left middle temporal gyrus to identify the face state, activity of upper parts of the inferior and middle temporal gyrus and parts of the fusiform gyrus to identify the place state, and

activity of the fusiform gyrus and right occipital pole to identify the tool state. Similar to 2D-DeepLight, it also associated the activity of the temporoparietal junction with the place state.

In summary, both DeepLight architectures (2D and 3D; see section 2.3.3) associated a similar overall set of brain regions with the four cognitive states of the HCP working memory task. These regions were also generally in line with a GLM analysis of the underlying fMRI data. Interestingly, the two architectures used the activity of different combinations of the overall set of brain regions to identify each of the four cognitive states.

Transfer to an independent fMRI dataset

Our previous analyses have demonstrated that a DeepLight variant that is pre-trained on the fMRI data of six out of the seven HCP experiment tasks outperforms a variant that is trained from scratch when both are applied to the fMRI data of the left-out seventh HCP experiment task. To test whether the pre-trained models exhibit a similar advantage when applied to an independent fMRI dataset that is not part of the HCP, we analyzed a dataset that was originally published by by Nakai and Nishimoto (the "Multi-task" dataset; Nakai and Nishimoto, 2020). In this dataset, six participants repeatedly performed 103 simple naturalistic tasks in the fMRI (e.g., deciding whether the music that is currently being played is Jazz or whether there is a penguin on a presented image; for further details on the tasks and dataset, see section 2.3.1 and Nakai and Nishimoto, 2020). The Multi-task dataset contains the fMRI data of 18 runs for each individual and is split into a training dataset (containing the data of 12 runs) and a test dataset (containing the data of the remaining six runs). In the test runs, participants performed versions of the 103 tasks that were not included in the training runs (for example, by utilizing different music or images). Similar to our previous analyses, we evaluated the performance of a pre-trained and not pre-trained variant of each DeepLight architecture (2D and 3D; see section 2.3.3) in identifying the 103 tasks (i.e., cognitive states) from the fMRI data of this dataset.

The 3D-DeepLight variants again performed slightly better at identifying the 103 tasks from the fMRI data than their 2D counterparts (see Appendix Fig. B.6; for an overview of the training methods, see Appendix B.1.7). While the pre-trained 2D-DeepLight variant achieved a final decoding accuracy of 37.96% in the test runs of the Multi-task dataset, the pre-trained 3D-DeepLight variant achieved a final decoding accuracy of 45.77% (i.e., 7.81% better than 2D-DeepLight, $t(5) = 9.24, P = 0.00025$). Similarly, the 2D-DeepLight variant that was not pre-trained achieved a final decoding accuracy of 33.02%, while the respective 3D-DeepLight variant achieved a final decoding accuracy of 43.33% (i.e., 10.31% better than 2D-DeepLight, $t(5) = 11.52, P < 0.0001$). Due to the generally better performance of 3D-DeepLight, we continued all subsequent analyses of the Multi-task dataset with the 3D-DeepLight architecture.

Surprisingly, the pre-trained 3D-DeepLight variant achieved only marginally better decoding accuracies in the test runs of the Multi-task dataset than the 3D-DeepLight variant that was not pre-trained, while both exhibited very similar learning speeds (see Appendix Fig. B.6 B). After 100 training epochs, the pre-trained 3D-DeepLight variant

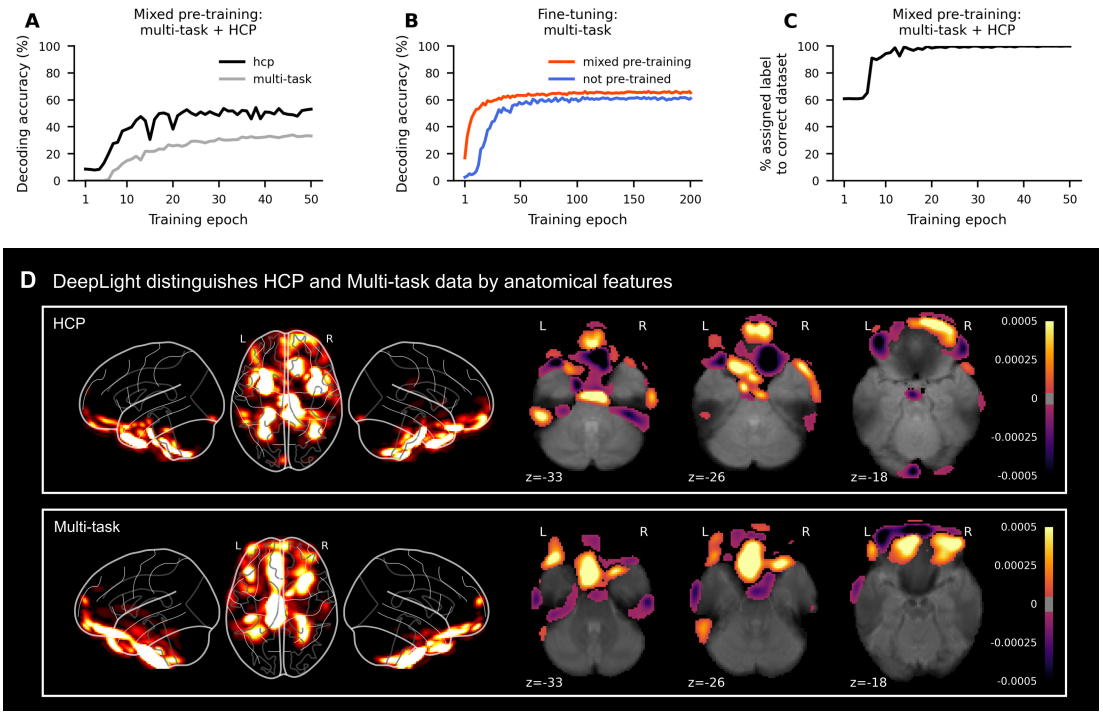

Figure 2.10: 3D-DeepLight learns separate mappings between brain activity and cognitive states when pre-trained with the data of the HCP and Multi-task datasets. A: During pre-training, 3D-DeepLight learns to accurately identify the cognitive states of both datasets (for an overview of the two datasets, see section 2.3.1). B: The pre-trained 3D-DeepLight variant (red) outperforms a model variant that is trained from scratch (blue), when both are applied to a left-out part of the Multi-task dataset. C: During pre-training, 3D-DeepLight learns to first separate the two datasets, by accurately assigning a cognitive state to each fMRI volume that is part of its dataset, while still learning to distinguish the cognitive states within each dataset (A). D: When training a 3D-DeepLight variant to solely distinguish the fMRI data of the HCP working memory task and Multi-task data, is uses anatomical features of the underlying brains to identify the each dataset (surrounding the brainstem, cerebellum, and ventromedial prefrontal cortex). Colors indicate average relevance values that results from an LRP decomposition of the model that is trained to separate the two fMRI datasets (for details on the LRP decomposition, see section 2.3.3). See the main text for further details on these analyses.

outperformed the variant that was not pre-trained by 2.44% ($t(5) = 6.40, P = 0.0014$).

To better understand whether the transfer performance of the pre-trained model to the Multi-task dataset was limited by the different preprocessing that we applied to the Multi-task and HCP data (we preprocessed the Multi-task dataset with fMRIPrep (Esteban et al., 2019), whereas the HCP uses an internal preprocessing pipeline; see section 2.3.1), we downloaded the raw fMRI data of another 50 subjects in the HCP working memory task and also preprocessed these with fMRIPrep (for an overview of the preprocessing steps, see Appendix B.1.12). Interestingly, the pre-trained 3D-DeepLight variant again exhibited the advantages of transfer learning in this newly preprocessed

fMRI dataset, by learning faster and achieving higher decoding accuracies than a model variant that was not pre-trained. After being trained on the fMRI data of 20 subjects from this newly preprocessed dataset, the pre-trained model achieved a final decoding accuracy of 72.95% in the fMRI data of the remaining 30 subjects, while the model variant that was not pre-trained achieved a final decoding accuracy of 64.46% (i.e., -8.49% worse than the pre-trained model, $t(29) = -13.28, P < 0.0001$; see Appendix Fig. B.7; for an overview of the training methods, see Appendix B.1.7).

We also tested whether basic differences in noise between the HCP and Multi-task datasets limited the transfer performance of the pre-trained model. To this end, we performed a confound correction of the Multi-task fMRI data, by regressing out variance related to the six motion correction parameters and three temporal and anatomical noise components resulting from fMRIPrep's CompCor method (for an overview, see Appendix Fig. B.8). However, the pre-trained model did not perform better when fine-tuned on the confound-corrected fMRI data than when fine-tuned on the fMRI data that was not confound corrected (for an overview of the training methods, see Appendix B.1.7). Its final decoding accuracy on the confound-corrected data was 43.27%, thereby -2.5% worse than when applied to the uncorrected fMRI data ($t(5) = -4.65, P = 0.0056$; Appendix Fig. B.8).

To test whether transfer learning can be at all beneficial for the Multi-task dataset, we pre-trained a 3D-DeepLight variant on a mixed dataset, composed of data from the HCP and Multi-task datasets (for an overview of the training methods, see Appendix B.1.7). We first assigned 70 randomly selected tasks of the Multi-task data to a pre-training dataset and the remaining 30 tasks to a test dataset. We then added the fMRI data of 50 randomly selected subjects in each of the six HCP pre-training tasks (see section 2.3.1 and 2.4.2) to the new pre-training dataset. We further divided the pre-training dataset into a distinct training and validation dataset, by assigning the data of the 12 training runs of the 70 pre-training tasks of the Multi-task data (see section 2.3.1), as well as the data of 45 randomly selected subjects of each HCP task, to a training dataset and the data of the remaining six test runs and HCP subjects to a test dataset. Note that the data of the 30 test tasks of the Multi-task dataset was similarly divided into a training and validation dataset, according to the 12 training and six test runs of this dataset.

After 50 training epochs, 3D-DeepLight accurately identified the cognitive states underlying 47.49% of the fMRI volumes in the validation data of the mixed pre-training dataset (56.70% within the HCP and 33.17% within the Multi-task data; Fig. 2.10 A). Subsequently, we compared the performance of this newly pre-trained 3D-DeepLight variant to a variant that was not pre-trained, when both are applied to the fMRI data of the remaining 30 tasks of the Multi-task dataset (for an overview of the training methods, see Appendix B.1.7). The newly pre-trained model again exhibited the advantages of transfer learning, by learning faster and achieving overall higher decoding accuracies than a variant that was not pre-trained. After 100 training epochs with the fMRI data of the training runs of the 30 Multi-task test tasks, the pre-trained model was able to accurately decode the cognitive state of 65.28% (Fig. 2.7 B) of the fMRI volumes of the test runs of the 30 test tasks, compared to 61.08% for the model variant that was trained from scratch (the pre-trained variant, by contrast, performed 4.2%

better; $t(5) = 4.62, P = 0.0057$; Fig. 2.7 B).

Interestingly, when analyzing the decoding decisions of the newly pre-trained 3D-DeepLight variant in more detail, we found that it treated both fMRI datasets separately (Fig. 2.10 C). Already early in its training, this variant learned to correctly identify the dataset underlying each fMRI volume, by accurately assigning one of the cognitive states of the respective dataset to the each volume, while still learning to distinguish the cognitive states within each dataset.

To better understand how 3D-DeepLight identified each of the two datasets, we trained a 3D-DeepLight variant to solely decode the dataset underlying each fMRI volume. For this analysis, we used the data of 50 subjects in the HCP working memory task, in addition to the full Multi-task dataset. We again subdivided these data into a distinct training and validation dataset, by assigning the fMRI data of 40 randomly selected HCP subjects and the training runs of the Multi-task data to a training dataset and the fMRI data of the remaining 10 subjects and test runs to a validation dataset. After 10 training epochs (for an overview of the training methods, see Appendix B.1.7), this 3D-DeepLight variant was able to perfectly identify (with 100% accuracy in the validation data) the dataset underlying each fMRI volume. To identify the features that this 3D-DeepLight variant used to distinguish between the two datasets, we decomposed its decoding decisions for the validation dataset with the LRP technique (for an overview of the LRP technique, see 2.3.3). Interestingly, it identified the two datasets solely by the use of the anatomical features of the underlying brains (surrounding the brainstem, cerebellum, and ventromedial prefrontal cortex; Fig. 2.10 D).

Based on these findings, we also tested whether the generalizability of a pre-trained model could be improved by restricting the pre-training data to only those brain regions whose activity is relevant to identify a set of cognitive states. To this end, we computed a brain mask for the fMRI data of the six HCP tasks that we use for pre-training (see section 2.3.1 and 2.4.2), by performing a random-effects GLM analysis for 50 randomly selected subjects in each of these tasks (for details on the GLM analysis, see Appendix B.1.8) and thresholding the resulting GLM brainmaps at a false-discovery rate of 0.1 (for an overview of the resulting brain mask, see Appendix Fig. B.9 A-B). After 16 training epochs on a masked version of the pre-training dataset (including the data of all 400 subjects in each of the six pre-training tasks; for an overview of the training methods, see Appendix B.1.7), 3D-DeepLight was able to correctly identify the cognitive states of 70.63% of the fMRI volumes in the masked validation data (see Appendix Fig. B.9 C). Yet, when fine-tuned on the fMRI data of the 12 training runs of the Multi-task dataset, it did not perform better than the variant that was pre-trained on the whole-brain fMRI data (see section 2.4.2). After 100 training epochs, the 3D-DeepLight variant that was trained on the masked pre-training data achieved a final decoding accuracy of 44.20% in the six test runs of the Multi-task dataset, while the model that was pre-trained on the whole-brain fMRI data achieved a decoding accuracy of 45.77% (see Appendix Fig. B.9 D).

In summary, these findings demonstrate that a DL model, which is pre-trained on fMRI data from the HCP, generalizes well to the fMRI data of other experiment tasks of the HCP, but not to an independent fMRI dataset that is not part of the HCP. Yet, when accounting for the independent and HCP data during pre-training, the pre-trained

model again exhibits the benefits of transfer learning, by learning faster and achieving higher decoding accuracies than a model variant that is trained from scratch, when both are applied to a left-out part of the independent dataset. Interestingly, the newly pre-trained model treats the two fMRI datasets separately, by learning to first perfectly identify which dataset each fMRI volume belongs to, while still learning to distinguish the cognitive states within each dataset.

2.5 Conclusion

The application of DL models to whole-brain fMRI data is generally hindered by two challenges. First, DL models act as "black boxes", disguising any relationship between a decoded cognitive state and the underlying brain activity. Second, conventional fMRI datasets are generally high-dimensional while containing comparably few samples. In this work, we investigated two solutions to these challenges, by the application of explanation techniques and transfer learning.

One of the main contributions of this work is the introduction of the DeepLight framework, which utilizes a DL model to decode a set of cognitive states from whole-brain fMRI data, and subsequently relates the decoded cognitive states and fMRI data, by interpreting the decoding decision with the LRP technique (Bach et al., 2015, Montavon et al., 2017).

In the following, we will summarize the findings of this work with respect to our research hypotheses (see section 2.2).

1. **A DL model can be trained to accurately decode a set of cognitive states from single whole-brain fMRI volumes.** Across two empirical studies, both DeepLight architectures (2D and 3D; see section 2.3.3) accurately decoded the cognitive states of six out of the seven HCP experiment tasks from single whole-brain fMRI volumes (with the exception of the gambling task; see Fig. 2.3 and Fig. 2.7). DeepLight was generally more accurate in decoding the four cognitive states underlying the HCP working memory task than the searchlight analysis and whole-brain lasso, on the level of the group (Fig. 2.3) and on the level of the individual (Fig. 2.4). Our findings therefore confirm this hypothesis.

2. **The application of the LRP technique to the decoding decisions of a DL model that accurately decodes a set of cognitive states from whole-brain fMRI data allows to identify a biologically plausible association between the decoded cognitive states and brain activity.** To test this hypothesis, we decomposed the decoding decisions of three DeepLight variants that accurately identified the four cognitive states of the HCP working memory task (see Fig. 2.5 and 2.9). In general, all three variants associated the activity of the same overall set of brain regions with the four cognitive states. This set was also generally in line with the results of a meta-analysis (Fig. 2.5 A) and a baseline GLM analysis of the same data (Fig. 2.5 B and Fig. 2.9 A). Our findings therefore generally confirm this hypothesis. Note, however, that the three DeepLight variants differed in the specific combination of the set of brain regions that they associated with each individual cognitive state. This indicates that there exist multiple valid solutions to this decoding problem, such that the activity of different combinations of brain regions can lead to a similar overall accuracy in distinguishing the four cognitive states.

3. **A DL model that is pre-trained on a large fMRI dataset achieves overall higher decoding accuracies and requires less training time and data than a model variant with the same architecture that is trained from**

scratch, when both are applied to the fMRI data of an *independent experiment task*. To test this hypothesis, we first trained both DeepLight architectures (2D and 3D; see section 2.3.3) on the fMRI data of 400 individuals in six of the seven HCP experiment tasks (all except for the working memory task; see Fig. 2.7). Subsequently, we compared the decoding performance of the pre-trained models to that of a respective DeepLight variant with the exact same architecture, which was trained from scratch, when they are applied to the fMRI data of the left-out HCP working memory task. The pre-trained DeepLight variants generally achieved higher decoding accuracies and required less training time and data than their not pre-trained counterparts. Our findings thereby confirm this hypothesis.

4. **A DL model that is pre-trained on a large fMRI dataset achieves overall higher decoding accuracies and requires less training time than a model variant with the same architecture that is trained from scratch, when both are applied to the fMRI data of an *independent fMRI dataset*.** To test this hypothesis, we compared the performance of both pre-trained DeepLight architectures (2D and 3D; see section 2.3.3) to a respective DeepLight variant with the exact same architecture that was not pre-trained, when they are applied to an fMRI dataset that is not part of the HCP. For this analysis, we used the "Multi-task" dataset, in which six human participants repeatedly performed 103 distinct simple tasks (for an overview of the dataset, see section 2.3.1 and Nakai and Nishimoto (2020)). In this analysis, the 3D-DeepLight architectures achieved generally better performance than their 2D-counterparts. Interestingly, the 3D-DeepLight architecture did not exhibit the advantages of pre-training that we previously observed in our analyses of the HCP data, by achieving only marginally better decoding accuracies than the variant that was not pre-trained and exhibiting very similar learning speeds (see Appendix Fig. B.6). To achieve the advantages of transfer learning, we needed to account for both datasets during pre-training, by dividing the data of the Multi-task dataset into two parts and including one part in a new pre-training dataset together with data from the HCP. Interestingly, the pre-trained model learned separate mappings between the fMRI data and cognitive states of the HCP and Multi-task datasets by first learning to identify the dataset underlying each fMRI volume, while still learning to accurately decode the cognitive states within each dataset. These findings thereby demonstrate that pre-training can in general be beneficial for the application of DL models to fMRI data, if the pre-training and fine-tuning data are representative of one another (for a more detailed discussion of this finding, see section 2.6.3). Our findings thereby, in general, confirm this hypothesis.

5. **The DeepLight framework is generalizable to different DL model architectures.** To test this hypothesis, we compared the performance of two DeepLight architectures (2D and 3D; see section 2.3.3) in our transfer learning analyses (see section 2.4.2 and 2.4). Both architectures accurately decoded the cognitive states from the fMRI data, while the 3D-DeepLight variants were slightly more accurate. The mapping between cognitive states and fMRI data that both archi-

tectures learned was generally also in line with the results of a meta-analysis and a random-effects GLM analysis of the same data (see Fig. 2.5 and 2.9. Interestingly, the specific combination of brain regions whose activity each architecture associated with each individual cognitive state was different between the two architectures. Our findings thereby confirm this hypothesis and demonstrate that the DeepLight framework is generalizable to the 2D- and 3D-DeepLight architectures.

In summary, this work has demonstrated that DL models, in combination with the LRP technique, provide a powerful framework for the analysis of fMRI data, by accurately decoding a set of cognitive states from whole-brain fMRI data and providing meaningful insights into the association of cognitive states and brain activity. Further, this work has demonstrated that the decoding performance of DL models in conventional fMRI datasets can generally be improved through transfer learning. Yet, future research is needed to identify the features of the fMRI data and the mapping between fMRI data and cognitive states that allow for transfer learning.

2.6 Discussion

2.6.1 The DeepLight framework and its relation to conventional analysis approaches for fMRI data

In the following, we will discuss the methodological differences between DeepLight and the three conventional analysis approaches for fMRI data that we considered in this work. DeepLight is defined by two central components (see Fig. 2.1): First, it utilizes a DL model to decode a cognitive state from whole-brain fMRI data. Subsequently, it relates the decoded cognitive state and fMRI data, by interpreting the decoding decision with the LRP technique. DL models are generally able to project the input data into a higher-level representation, in which they can associate a target signal with variable patterns in the data. While DL models are considered data-driven and can thereby autonomously learn these projections from the data, they also require sufficiently large training datasets to generalize well to new data.

General linear model: The GLM is conceptually different from the other analysis approaches that we considered. It aims to identify an association between cognitive state and brain activity by predicting the time series signal of a single voxel from a linear combination of a set of experiment predictors. The GLM thereby treats each voxel's signal as independent from others. Due to these simplifications, the mappings between fMRI data and cognitive states learned by the GLM are generally simpler than those of the other approaches, while it has a comparably little risk of overfitting. Yet, the performance of the GLM in predicting the response signal of a voxel is typically not evaluated out-of-sample, which generally leaves unanswered how well its results generalize to new data.

Searchlight analysis: DeepLight generally outperformed the searchlight analysis in decoding the cognitive states from the fMRI data. In small datasets (here, 10 or fewer subjects), however, the performance of the searchlight analysis was superior. The searchlight analysis decodes a cognitive state from a single cluster of only few voxels. Its input data, as well as the number of parameters in its decoding model, are thereby considerably smaller, leading to an overall lower risk of overfitting. Yet, this advantage comes at the cost of additional constraints, which have to be considered when considering both approaches. If a cognitive state is associated with the activity of a small brain region only, the searchlight analysis will generally be more sensitive to the activity of this region, as it has learned a decoding model that is specific to the activity of the region. If, however, the cognitive state is not identifiable by the activity of a single brain region, but solely in conjunction with the activity of another, spatially distinct, brain region, the searchlight analysis will not be able to identify this association due to its narrow spatial focus. DeepLight, on the other hand, will generally be less sensitive to the specifics of the activity of a local brain region, but perform better in identifying a cognitive state from spatially distributed patterns of brain activity. When choosing between the two approaches, one should therefore consider whether the assumed associations between brain activity and cognitive state involve a set of spatially local or

distributed brain regions, and whether the amount of available fMRI data allows for the application of whole-brain decoding approaches.

Whole-brain least absolute shrinkage regression: The whole-brain lasso is based on a linear decoding model. It assigns a single coefficient weight to each voxel in the brain and makes a decoding decision by computing a weighted sum over the activity of an input fMRI volume. Importantly, due to the strong regularization that is applied to the coefficients during the training, many coefficients equal 0. The resulting set of coefficients therefore resembles a brain mask that defines a fixed set of brain regions whose activity the whole-brain lasso utilizes to decode a cognitive state. DeepLight, on the other hand, makes a decoding decision by sequentially projecting the input fMRI data into a higher-level representation. This abstracted (and more flexible) view enables DeepLight to better account for the variable patterns of brain activity underlying a cognitive state (within and across individuals). This ability is exemplified in Figure 2.6, as well as Supplementary Videos 1, 2 of Thomas et al. (2019a) wherein we visualize the variable patterns of brain activity that DeepLight associates with the face and place stimulus classes throughout an experiment block. The relevance patterns of DeepLight mimic the hemodynamic response and peak in the ROIs 5–10 s after the onset of the experiment block. Importantly, we find that the whole-brain lasso does not exhibit such temporo-spatial variability. Further, in our analyses with the fMRI data of 100 individuals in the HCP working memory task, DeepLight generally outperformed the whole-brain lasso in an out-of-sample decoding prediction. DeepLight was thereby able to learn a mapping between fMRI data and cognitive states that generalized better to a test dataset than the mapping learned by the whole-brain lasso.

2.6.2 Limitations of the DeepLight framework

We have demonstrated that DeepLight is able to accurately decode a set of cognitive states from whole-brain fMRI data and identify a biologically plausible association between these cognitive states and the fMRI data. However, our findings have also pointed at several limitations of the DeepLight framework, which we will discuss in the following.

First, similar to other DL models, DeepLight requires a comparably large training dataset to generalize well to new data. In our analyses of the HCP working memory task, DeepLight required the fMRI data of at least 50 individuals to outperform the searchlight analysis and whole-brain lasso in the out-of-sample decoding prediction. DeepLight is therefore not well suited for datasets with only few subjects. This is also exemplified in Fig. 2.8 A, D and Fig. 2.3 E, where DeepLight struggles to identify an association between cognitive states and fMRI data that generalizes well to the test datasets, when being trained with the data of less than 10 individuals.

Second, when interpreting the decoding decisions of three DeepLight variants that accurately decoded the cognitive states of the HCP working memory task (see Fig. 2.5 E and 2.9 B-C), we found that they all generally utilized activity from the same overall set of brain regions to distinguish the four cognitive states. Yet, they associated the activity of different combinations of these regions with each individual cognitive state. This

indicates that there exist multiple solutions to this decoding problem that result in a similar overall decoding performance. Future applications of the DeepLight framework should therefore control for the effects of convergence on different local optima during training. Ideally, by training the same model in multiple runs on the same dataset to approximate any variability in the learned mapping between cognitive states and fMRI data that results from convergence on different local optima. Similarly, DeepLight's learned mappings between cognitive states and fMRI data should generally be interpreted in relation to other empirical findings or the results of other analysis approaches. In our analyses, performing a random-effects GLM analysis of the same fMRI data allowed us to identify that 3D-DeepLight learned to decode each cognitive state of the HCP working memory task through the activity of those brain regions that generally exhibit less activity in this cognitive state than in the others (see Fig. 2.9 C).

Third, as for other decoding approaches, DeepLight's learned mappings between cognitive states and fMRI data are generally dependent on the given decoding problem and dataset. In theory, it is possible for any decoding model to accurately identify all cognitive states of a dataset by learning to correctly distinguish all but one cognitive state from the others. In this case, all samples that belong to neither of these cognitive states can be correctly classified as the remaining cognitive state, without any knowledge about the actual mapping between the fMRI data and the remaining cognitive state. Any interpretation of DeepLight's learned mapping between cognitive states and fMRI data should therefore always consider the underlying decoding task and data.

2.6.3 Transfer learning with fMRI data

A DeepLight variant that is pre-trained on the fMRI data of six HCP experiment tasks generally outperforms a variant with the exact same architecture that is trained from scratch, when both are applied to the data of the left-out seventh HCP experiment task. Yet, learning about the mapping between cognitive states and fMRI data of the HCP experiment tasks did not benefit DeepLight's application to an fMRI dataset that is not part of the HCP. This finding is surprising, as ML research indicates that transfer learning, even between distant tasks, is generally beneficial compared to random weight initialization (Yosinski et al., 2014).

Interestingly, when training a DeepLight variant on a dataset that combined fMRI data from the HCP and independent dataset, we found that it learned a separate mapping between the fMRI data and cognitive states of both datasets, by learning to first identify to which dataset each fMRI volume belonged, while still learning to distinguish the cognitive states within each dataset. Note that this was not the case for our analysis within the HCP data (see Fig. 2.7 D, H), as DeepLight there still exhibited some confusion between the cognitive states of different HCP experiment tasks (in spite of its overall higher decoding accuracy).

Surprisingly, the transfer of the pre-trained model to the independent fMRI dataset was neither hindered by the differences in preprocessing between the two datasets, nor by basic differences in noise. We can further rule out differences in the temporal distribution of brain activity between the two datasets, as DeepLight processes single fMRI volumes independent from one another, and we detrended and standardized the

time-series signal of each voxel in each run (see section 2.3.1), such that all voxels had very similar basic temporal signal characteristics.

These findings therefore have important implications for the future study of transfer learning with fMRI data, as they raise the question which features of the mapping between fMRI data and cognitive states are specific to a dataset and which generalize to others. In addition, future research is needed to explore whether there is a boundary condition under which a DL model stops treating different fMRI datasets separately and starts learning features of the mapping between brain activity and cognitive states that generalize well to other datasets.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *12th Symposium on Operating Systems Design and Implementation*, pages 265–283.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8(14).

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, 12(2):169–177.

Alber, M., Lapuschkin, S., Seegerer, P., Hagele, M., Schutt, K. T., Montavon, G., Samek, W., Muller, K.-R., Dahne, S., and Kindermans, P.-J. (2019). iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20:1–8.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L., and Smith, S. M. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424.

Alós-Ferrer, C. (2018). A Dual-Process Diffusion Model. *Journal of Behavioral Decision Making*, 31(2):203–218.

Armel, K. C., Beaumel, A., and Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making*, 3(5):396–403.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. (2017a). "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):e0181142.

Arras, L., Montavon, G., Müller, K.-R., and Samek, W. (2017b). Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Ashby, N. J. S., Jekel, M., Dickert, S., and Glöckner, A. (2016). Finding the right fit: A comparison of process assumptions underlying popular drift-diffusion models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(12):1982–1993.

Avants, B., Epstein, C., Grossman, M., and Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Babayan, A., Erbey, M., Kumral, D., Reinelt, J. D., Reiter, A. M. F., Röbbig, J., Schaare, H. L., Uhlig, M., Anwander, A., Bazin, P.-L., Horstmann, A., Lampe, L., Nikulin, V. V., Okon-Singer, H., Preusser, S., Pampel, A., Rohr, C. S., Sacher, J., Thöne-Otto, A., Trapp, S., Nierhaus, T., Altmann, D., Arelin, K., Blöchl, M., Bongartz, E., Breig, P., Cesnaite, E., Chen, S., Cozatl, R., Czerwonatis, S., Dambrauskaite, G., Dreyer, M., Enders, J., Engelhardt, M., Fischer, M. M., Forschack, N., Golchert, J., Golz, L., Guran, C. A., Hedrich, S., Hentschel, N., Hoffmann, D. I., Huntenburg, J. M., Jost, R., Kosatschek, A., Kunzendorf, S., Lammers, H., Lauckner, M. E., Mahjoory, K., Kanaan, A. S., Mendes, N., Menger, R., Morino, E., Näthe, K., Neubauer, J., Noyan, H., Oligschläger, S., Panczyszyn-Trzewik, P., Poehlchen, D., Putzke, N., Roski, S., Schaller, M.-C., Schieferbein, A., Schlaak, B., Schmidt, R., Gorgolewski, K. J., Schmidt, H. M., Schrimpf, A., Stasch, S., Voss, M., Wiedemann, A., Margulies, D. S., Gaebler, M., and Villringer, A. (2019). A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6(1):180308.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., and Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189.

Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Beer, J. C., Aizenstein, H. J., Anderson, S. J., and Krafty, R. T. (2019). Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages. *Biometrics*, 75(4):1299–1309.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007a). A component based noise correction method (CompCor) for BOLD and perfusion based fmri. *NeuroImage*, 37(1):90–101.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007b). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1):90–101.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Berkowitsch, N. A. J., Scheibehenne, B., and Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3):1331–1348.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, 120(3):522–543.

Boorman, E. D., Rushworth, M. F., and Behrens, T. E. (2013). Ventromedial Prefrontal and Anterior Cingulate Cortex Adopt Choice and Default Reference Frames during Sequential Multi-Alternative Choice. *Journal of Neuroscience*, 33(6):2242–2253.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv:1508.05326 [cs]*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Callaway, F., Rangel, A., and Griffiths, T. (2019). Fixation patterns in simple choice are consistent with optimal use of cognitive resources. Preprint, PsyArXiv.

Caplin, A., Dean, M., and Martin, D. (2011). Search and Satisficing. *American Economic Review*, 101(7):2899–2922.

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., Chaarani, B., Mejia, M. H., Hagler, D. J., Daniela Cornejo, M., Sicat, C. S., Harms, M. P., Dosenbach, N. U. F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J. R., Kuperman, J. M., Fair, D. A., and Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., and Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143(4):1476–1488.

Chandon, P., Hutchinson, J. W., Bradlow, E. T., and Young, S. H. (2009). Does In-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase. *Journal of Marketing*, 73(6):1–17.

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., and Wager, T. D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biology*, 13(6):e1002180.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Clithero, J. A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cox, R. W. and Hyde, J. S. (1997). Software tools for analysis and visualization of fmri data. *NMR in Biomedicine*, 10(4-5):171–178.

Cremers, H. R., Wager, T. D., and Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLOS ONE*, 12(11):e0184923.

Decety, J. and Grèzes, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research*, 1079(1):4–14.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Diederich, A. (2003). MDFT account of decision making under time pressure. *Psychonomic Bulletin & Review*, 10(1):157–166.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634.

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, 293(5539):2470–2473.

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Kent, James D. andGoncalves, M., DuPre, E., Sitek, K. R., Gomez, D. E. P., Lurie, D. J., Ye, Z., Poldrack, R. A., and Gorgolewski, K. J. (2018). fmriprep. *Software*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116.

Evans, A., Janke, A., Collins, D., and Baillet, S. (2012). Brain templates and atlases. *NeuroImage*, 62(2):911–922.

Fehr, E. and Rangel, A. (2011). Neuroeconomic Foundations of Economic Choice–Recent Advances. *Journal of Economic Perspectives*, 25(4):3–30.

Fellner, G., Güth, W., and Maciejovsky, B. (2009). Satisficing in financial decision making — a theoretical and experimental approach to bounded rationality. *Journal of Mathematical Psychology*, 53(1):26–33.

Fellows, L. K. (2006). Deciding how to decide: Ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. *Brain*, 129(4):944–952.

Fiedler, S. and Glöckner, A. (2012). The Dynamics of Decision Making in Risky Choice: An Eye-Tracking Analysis. *Frontiers in Psychology*, 3.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2):774–781.

Fisher, G. (2017). An attentional drift diffusion model over binary-attribute choice. *Cognition*, 168:34–45.

Folke, T., Jacobsen, C., Fleming, S. M., and Martino, B. D. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1):1–8.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1:S102.

Fox, C. J., Iaria, G., and Barton, J. J. S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. *Human Brain Mapping*, 30(5):1637–1651.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, 24(1):244–252.

Gazzaniga, M. S. (2006). *Handbook of Functional Neuroimaging of Cognition*. MIT Press.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, 15(4):870–878.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

Glimcher, P. W. and Fehr, E. (2013). *Neuroeconomics: Decision Making and the Brain*. Academic Press.

Glöckner, A. and Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, 24(1):71–98.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gluth, S., Kern, N., Kortmann, M., and Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, 4(6):634–645.

Gluth, S., Spektor, M. S., and Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife*, 7:e39659.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, 5.

Gorgolewski, K. J., Esteban, O., Markiewicz, C. J., Ziegler, E., Ellis, D. G., Notter, M. P., Jarecka, D., Johnson, H., Burns, C., Manhães-Savio, A., Hamalainen, C., Yvernault, B., Salo, T., Jordan, K., Goncalves, M., Waskom, M., Clark, D., Wong, J., Loney, F., Modat, M., Dewey, B. E., Madison, C., Visconti di Oleggio Castello, M., Clark, M. G., Dayan, M., Clark, D., Keshavan, A., Pinsard, B., Gramfort, A., Berleant, S., Nielson, D. M., Bougacha, S., Varoquaux, G., Cipollini, B., Markello, R., Rokem, A., Moloney, B., Halchenko, Y. O., Wassermann, D., Hanke, M., Horea, C., Kaczmarzyk, J., de Hollander, G., DuPre, E., Gillman, A., Mordom, D., Buchanan, C., Tungaraza, R., Pauli, W. M., Iqbal, S., Sikka, S., Mancini, M., Schwartz, Y., Malone, I. B., Dubois, M., Frohlich, C., Welch, D., Forbes, J., Kent, J., Watanabe, A., Cumba, C., Huntenburg, J. M., Kastman, E., Nichols, B. N., Eshaghi, A., Ginsburg, D., Schaefer, A., Acland, B., Giavasis, S., Kleesiek, J., Erickson, D., Küttner, R., Haselgrove, C., Correa, C., Ghayoor, A., Liem, F., Millman, J., Haehn, D., Lai, J., Zhou, D., Blair, R., Glatard, T., Renfro, M., Liu, S., Kahn, A. E., Pérez-García,

F., Triplett, W., Lampe, L., Stadler, J., Kong, X.-Z., Hallquist, M., Chetverikov, A., Salvatore, J., Park, A., Poldrack, R., Craddock, R. C., Inati, S., Hinds, O., Cooper, G., Perkins, L. N., Marina, A., Mattfeld, A., Noel, M., Snoek, L., Matsubara, K., Cheung, B., Rothmei, S., Urchs, S., Durnez, J., Mertz, F., Geisler, D., Floren, A., Gerhard, S., Sharp, P., Molina-Romero, M., Weinstein, A., Broderick, W., Saase, V., Andberg, S. K., Harms, R., Schlamp, K., Arias, J., Papadopoulos Orfanos, D., Tarbert, C., Tambini, A., De La Vega, A., Nickson, T., Brett, M., Falkiewicz, M., Podranski, K., Linkersdörfer, J., Flandin, G., Ort, E., Shachnev, D., McNamee, D., Davison, A., Varada, J., Schwabacher, I., Pellman, J., Perez-Guevara, M., Khanuja, R., Pannetier, N., McDermottroe, C., and Ghosh, S. (2018). Nipype. *Software*.

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). Identifying Predictive Regions from fMRI with TV-L1 Prior. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 17–20.

Grandy, T. H., Lindenberger, U., and Werkle-Bergner, M. (2017). When Group Means Fail: Can One Size Fit All? *bioRxiv*.

Greve, D. N. and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., and Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72:304–321.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. (2019). SpotTune: Transfer Learning Through Adaptive Fine-Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4805–4814.

Guthery, F. S., Burnham, K. P., and Anderson, D. R. (2003). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. *The Journal of Wildlife Management*, 67(3):655.

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.-R., and Binder, A. (2020). Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific Reports*, 10(1):6423.

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science*, 324(5927):646–648.

Hauser, J. R. and Wernerfelt, B. (1990). An Evaluation Cost Model of Consideration Sets. *Journal of Consumer Research*, 16(4):393–408.

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

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233.

Hayes, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review*, 60(4):269–275.

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*.

Hecht-Nielsen, R. (1992). Theory of the Backpropagation Neural Network. In Wechsler, H., editor, *Neural Networks for Perception*, pages 65–93. Academic Press.

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, 431(7010):859–862.

Helfinstein, S. M., Schonberg, T., Congdon, E., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Cannon, T. D., London, E. D., Bilder, R. M., and Poldrack, R. A. (2014). Predicting risky choices from brain activity patterns. *Proceedings of the National Academy of Sciences*, 111(7):2470–2475.

Hinterstoisser, S., Lepetit, V., Wohlhart, P., and Konolige, K. (2018). On Pre-Trained Image Features and Synthetic Images for Deep Learning. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 1–15.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Holmes, A. P. and Friston, K. J. (1998). Generalisability, Random Effects & Population Inference. *NeuroImage*, 7(4, Part 2):S754.

Horst, F., Lapuschkin, S., Samek, W., Müller, K.-R., and Schöllhorn, W. I. (2019). Explaining the unique nature of individual gait patterns with deep learning. *Scientific Reports*, 9(1):2391.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). Functional Magnetic Resonance Imaging, Second Edition.

Hunt, L. T., Malalasekera, W. M. N., de Berker, A. O., Miranda, B., Farmer, S. F., Behrens, T. E. J., and Kennerley, S. W. (2018). Triple dissociation of attention and decision computations across prefrontal cortex. *Nature Neuroscience*, 21(10):1471–1481.

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron*, 87(2):451–462.

Jang, A., Sharma, R., and Drugowitsch, J. (2020). Optimal policy for attention-modulated decisions explains human fixation behavior. *bioRxiv*.

Jang, H., Plis, S. M., Calhoun, V. D., and Lee, J.-H. (2017). Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. *NeuroImage*, 145:314–328.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156.

Kauffmann, J., Müller, K.-R., and Montavon, G. (2020). Towards explaining anomalies: A deep Taylor decomposition of one-class models. *Pattern Recognition*, 101:107198.

Khan, S., Islam, N., Jan, Z., Ud Din, I., and Rodrigues, J. J. P. C. (2019). A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic Estimation of the Maximum of a Regression Function. *Annals of Mathematical Statistics*, 23(3):462–466.

Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., and Lapuschkin, S. (2019). Towards best practice in explaining neural network decisions with LRP. *arXiv:1910.09840 [cs, stat]*.

Konovalov, A. and Krajbich, I. (2016). Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nature Communications*, 7(1):12438.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., and Fehr, E. (2015). A Common Mechanism Underlying Food Choice and Social Decisions. *PLOS Computational Biology*, 11(10):e1004371.

Krajbich, I., Lu, D., Camerer, C., and Rangel, A. (2012). The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions. *Frontiers in Psychology*, 3.

Krajbich, I. and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857.

Kriegeskorte, N. and Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, 55:167–179.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Krogh, A. and Hertz, J. A. (1992). A Simple Weight Decay Can Improve Generalization. In Moody, J. E., Hanson, S. J., and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems 4*, pages 950–957. Morgan-Kaufmann.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Lanczos, C. (1964). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):76–85.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lewandowsky, S. and Farrell, S. (2010). *Computational Modeling in Cognition: Principles and Practice*. SAGE Publications.

Lindquist, M. A., Meng Loh, J., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage*, 45(1, Supplement 1):S187–S198.

Lopez-Persem, A., Domenech, P., and Pessiglione, M. (2016). How prior preferences determine decision-making frames and biases in the human brain. *eLife*, 5:e20317.

Loula, J., Varoquaux, G., and Thirion, B. (2018). Decoding fMRI activity in the time domain improves classification performance. *NeuroImage*, 180:203–210.

Luce, R. D. and Raiffa, H. (1957). Games & Decisions. John Willey & Sons. *Inc., New York*.

Mahmood, U., Rahman, M. M., Fedorov, A., Fu, Z., Calhoun, V. D., and Plis, S. M. (2019). Learnt dynamics generalizes across tasks, datasets, and populations. *arXiv:1912.03130 [cs]*.

Marban, A., Srinivasan, V., Samek, W., Fernández, J., and Casals, A. (2019). A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery. *Biomedical Signal Processing and Control*, 50:134–150.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Feczko, E., Dominguez, O. M., Graham, A., Earl, E. A., Perrone, A., Cordova, M., Doyle, O., Moore, L. A., Conan, G., Uriarte, J., Snider, K., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Laumann, T. O., Thompson, W. K., Greene, D. J., Petersen, S. E., Nichols, T., Yeo, B. T. T., Barch, D. M., Garavan, H., Luna, B., Fair, D. A., and Dosenbach, N. U. F. (2020). Towards Reproducible Brain-Wide Association Studies. *bioRxiv*.

Martino, B. D., Kumaran, D., Seymour, B., and Dolan, R. J. (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. *Science*, 313(5787):684–687.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed Attention. *American Economic Review*, 102(5):2183–2205.

Matějka, F. and McKay, A. (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–298.

McCall, J. J. (1970). Economics of Information and Job Search. *The Quarterly Journal of Economics*, 84(1):113–126.

McIntosh, A. R. and Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage*, 23:S250–S263.

Mensch, A., Mairal, J., Thirion, B., and Varoquaux, G. (2018). Extracting Universal Representations of Cognition across Brain-Imaging Studies. *arXiv:1809.06035 [cs, q-bio, stat]*.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total Variation Regularization for fMRI-Based Prediction of Behavior. *IEEE Transactions on Medical Imaging*, 30(7):1328–1340.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6):437–449.

Milosavljevic, M., Navalpakkam, V., Koch, C., and Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, 22(1):67–74.

Molter, F., Thomas, A. W., Heekeren, H. R., and Mohr, P. N. C. (2019). GLAMbox: A Python toolbox for investigating the association between gaze allocation and decision behaviour. *PLOS ONE*, 14(12):e0226428.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Morioka, H., Calhoun, V., and Hyvärinen, A. (2020). Nonlinear ICA of fMRI reveals primitive temporal structures linked to rest, task, and behavioral traits. *NeuroImage*, 218:116989.

Müller, K.-R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Murphy, K., Bodurka, J., and Bandettini, P. A. (2007). How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *NeuroImage*, 34(2):565–574.

Nakai, T. and Nishimoto, S. (2020). Quantitative models reveal the organization of diverse cognitive functions in the brain. *Nature Communications*, 11(1):1142.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 78, New York, NY, USA. Association for Computing Machinery.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016a). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3387–3395. Curran Associates, Inc.

Nguyen, A., Yosinski, J., and Clune, J. (2016b). Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. *arXiv:1602.03616 [cs]*.

Noguchi, T. and Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1):44–56.

Olson, I. R., Plotzker, A., and Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain*, 130(7):1718–1731.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., and Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, 112(13):4170–4175.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1310–III–1318, Atlanta, GA, USA. JMLR.org.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, 16(2):366–387.

Peelen, M. V. and Downing, P. E. (2005). Within-subject reproducibility of category-specific visual activation with functional MRI. *Human Brain Mapping*, 25(4):402–408.

Philiastides, M. G. and Ratcliff, R. (2013). Influence of Branding on Preference-Based Decision Making. *Psychological Science*, 24(7):1208–1215.

Pieters, R. and Warlop, L. (1999). Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, 16(1):1–16.

Pinar Saygin, A., Cicekli, I., and Akman, V. (2000). Turing Test: 50 Years Later. *Minds and Machines*, 10(4):463–518.

Polanía, R., Woodford, M., and Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, 22(1):134–142.

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M. (2013). Toward open sharing of task-based fMRI data: The OpenfMRI project. *Frontiers in Neuroinformatics*, 7.

Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Freimer, N. B., London, E. D., Cannon, T. D., and Bilder, R. M. (2016). A phenome-wide examination of neural and cognitive function. *Scientific Data*, 3(1):160110.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, 84(Supplement C):320–341.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753):477–486.

Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S., and Thoma, G. R. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, 6:e4568.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*.

Rapoport, A. and Tversky, A. (1966). Cost and accessibility of offers as determinants of optional stopping. *Psychonomic Science*, 4(1):145–146.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2):59–108.

Ratcliff, R. and McKoon, G. (2007). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, 20(4):873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281.

Ratcliff, R., Thapar, A., and McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, 13(4):626–635.

Ratcliff, R., Thapar, A., and McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, 60(3):127–157.

Ratcliff, R. and Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3):438–481.

Reutskaja, E., Nagel, R., Camerer, C. F., and Rangel, A. (2011). Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study. *American Economic Review*, 101(2):900–926.

Reverberi, C., Kuhlen, A. K., Seyed-Allaei, S., Greulich, R. S., Costa, A., Abutalebi, J., and Haynes, J.-D. (2018). The neural basis of free language choice in bilingual speakers: Disentangling language choice and language execution. *NeuroImage*, 177:108–116.

Roberts, J. H. and Lattin, J. M. (1991). Development and Testing of a Model of Consideration Set Composition. *Journal of Marketing Research*, 28(4):429–440.

Rodriguez, C. A., Turner, B. M., and McClure, S. M. (2014). Intertemporal Choice as Discounted Value Accumulation. *PLOS ONE*, 9(2):e90138.

Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, 108(2):370–392.

Rosa, M. J., Portugal, L., Hahn, T., Fallgatter, A. J., Garrido, M. I., Shawe-Taylor, J., and Mourao-Miranda, J. (2015). Sparse network-based models for patient classification using fMRI. *NeuroImage*, 105:493–506.

Russo, J. E. and Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, 3(3):267–276.

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2):752–764.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Samek, W., Binder, A., Lapuschkin, S., and Müller, K.-R. (2017). Understanding and Comparing Deep Neural Networks for Age and Gender Classification. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1629–1638, Venice. IEEE.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2020). Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv:2003.07631 [cs, stat]*.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64(1):240–256.

Schmiedek, F., Lövdén, M., and Lindenberger, U. (2009). On the relation of mean reaction time and intraindividual reaction time variability. *Psychology and Aging*, 24(4):841–857.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. (2020). XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv:2006.03589 [cs, stat]*.

Schölkopf, B., Smola, A. J., and Bach, F. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schuck, N. W., Cai, M. B., Wilson, R. C., and Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, 91(6):1402–1412.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. (2018). SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178–1197.

Shimojo, S., Simion, C., Shimojo, E., and Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138.

Simon, H. A. (1957). *Models of Man; Social and Rational*. Models of Man; Social and Rational. Wiley, Oxford, England.

Simon, H. A. (1959). Theories of Decision-Making in Economics and Behavioral Science. *The American Economic Review*, 49(3):253–283.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690.

Smith, S. M. and Krajbich, I. (2018). Attention and choice across domains. *Journal of Experimental Psychology: General*, 147(12):1810–1826.

Smith, S. M. and Krajbich, I. (2019). Gaze Amplifies Value in Decision Making. *Psychological Science*, 30(1):116–128.

Smith, S. M., Krajbich, I., and Webb, R. (2019). Estimating the dynamic role of attention via random utility. *Journal of the Economic Science Association*, 5(1):97–111.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stewart, N., Gächter, S., Noguchi, T., and Mullett, T. L. (2016a). Eye Movements in Strategic Choice. *Journal of Behavioral Decision Making*, 29(2-3):137–156.

Stewart, N., Hermens, F., and Matthews, W. J. (2016b). Eye Movements in Risky Choice. *Journal of Behavioral Decision Making*, 29(2-3):116–136.

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods*, 274:141–145.

Stüttgen, P., Boatwright, P., and Monroe, R. T. (2012). A Satisficing Choice Model. *Marketing Science*, 31(6):878–899.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Tavares, G., Perona, P., and Rangel, A. (2017). The Attentional Drift Diffusion Model of Simple Perceptual Decision-Making. *Frontiers in Neuroscience*, 11.

Thomas, A., Molter, F., and Krajbich, I. (2020). Uncovering the Computational Mechanisms Underlying Many-Alternative Choice. Preprint, PsyArXiv.

Thomas, A. W., Heekeren, H. R., Müller, K.-R., and Samek, W. (2019a). Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. *Frontiers in Neuroscience*, 13.

Thomas, A. W., Molter, F., Krajbich, I., Heekeren, H. R., and Mohr, P. N. C. (2019b). Gaze bias differences capture individual choice behaviour. *Nature Human Behaviour*, 3(6):625–635.

Thomas, A. W., Müller, K.-R., and Samek, W. (2019c). Deep Transfer Learning for Whole-Brain FMRI Analyses. In Zhou, L., Sarikaya, D., Kia, S. M., Speidel, S., Malpani, A., Hashimoto, D., Habes, M., Löfstedt, T., Ritter, K., and Wang, H., editors, *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, Lecture Notes in Computer Science, pages 59–67, Cham. Springer International Publishing.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., and Vuust, P. (2014). Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data. *NeuroImage*, 88:170–180.

Towal, R. B., Mormann, M., and Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, 110(40):E3858.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning Spatiotemporal Features With 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497.

Trueblood, J. S., Brown, S. D., and Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2):179–205.

Trueblood, J. S., Brown, S. D., Heathcote, A., and Busemeyer, J. R. (2013). Not Just for Consumers: Context Effects Are Fundamental to Decision Making. *Psychological Science*, 24(6):901–908.

Tsetsos, K., Chater, N., and Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, 109(24):9659–9664.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Uğurbil, K., Xu, J., Auerbach, E. J., Moeller, S., Vu, A. T., Duarte-Carvajalino, J. M., Lenglet, C., Wu, X., Schmitter, S., Van de Moortele, P. F., Strupp, J., Sapiro, G., De Martino, F., Wang, D., Harel, N., Garwood, M., Chen, L., Feinberg, D. A., Smith, S. M., Miller, K. L., Sotiropoulos, S. N., Jbabdi, S., Andersson, J. L. R., Behrens, T. E. J., Glasser, M. F., Van Essen, D. C., and Yacoub, E. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *NeuroImage*, 80:80–104.

Usher, M. and McClelland, J. L. (2004). Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice. *Psychological Review*, 111(3):757–769.

Vaidya, A. R. and Fellows, L. K. (2015). Testing necessary regional frontal contributions to value assessment and fixation-based updating. *Nature Communications*, 6(1):10120.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79.

Varoquaux, G. and Craddock, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject Dictionary Learning to Segment an Atlas of Brain Spontaneous Activity. In Székely, G. and Hahn, H. K., editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 562–573, Berlin, Heidelberg. Springer.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

von Neumann, J. and Morgenstern, O. (2007). *Theory of Games and Economic Behavior (Commemorative Edition)*. Princeton University Press.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine*, 368(15):1388–1397.

Webb, R. (2018). The (Neural) Dynamics of Stochastic Choice. *Management Science*, 65(1):230–255.

Westbrook, A., van den Bosch, R., Määttä, J. I., Hofmans, L., Papadopetraki, D., Cools, R., and Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, 367(6484):1362–1366.

Westfall, J. (2017). Statistical details of the default priors in the Bambi library. *arXiv:1702.01201 [stat]*.

Wiecki, T. V., Sofer, I., and Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, 7.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665–670.

Yarkoni, T. and Westfall, J. (2016). Bambi: A simple interface for fitting Bayesian mixed effects models. *OSF*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc.

Zhang, H., Chen, P.-H., and Ramadge, P. (2018). Transfer Learning on fMRI Datasets. In *International Conference on Artificial Intelligence and Statistics*, pages 595–603. PMLR.

Zhang, Y. and Bellec, P. (2020). Transferability of Brain decoding using Graph Convolutional Networks. *bioRxiv*.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57.

Appendix A

Appendix: The role of gaze allocation in simple choice

A.1 Methodology

A.1.1 Dataset details

Krajbich 2010 This dataset contains choice and eye-tracking data from a prototypical value-based, two-item snack food choice experiment. It includes data from 39 Caltech students, who reported to regularly eat the snack foods used in the experiment and had no dietary restrictions. They received a $20 show-up fee and one food item. All participants were asked not to eat for 3 hours prior to the experiment. In an initial liking rating task participants indicated liking ratings between -10 to 10 for each of 70 different snack food items (e.g., potato chips, candy bars) using an on screen slider with a randomized starting point and no time restriction ("How much would you like to eat this at the end of the experiment?"). These ratings were used as a measure of the value participants placed on each item. In the subsequent choice task participants made choices between pairs of food items. Choices were indicated without time restrictions using the left and right arrow keys on a keyboard. Each trial began with a 2 s forced fixation towards the center of the screen. A yellow feedback box was shown around the chosen item for 1 s after a choice was made. After the choice task, participants were required to stay for additional 30 min, to eat a food item that they chose in one randomly selected choice trial. Each participant performed 100 choice trials. Eye movements were continuously recorded in the choice task, using a 50 Hz desktop-mounted Tobii eye tracker. The data ere obtained from the original authors in a preprocessed format. The original preprocessing steps included the removal of trials with missing fixation data for more than 40 ms at the beginning or end of the trial (mean $\pm$ s.e.m. number of trials dropped per participant was 2.8 $\pm$ 1.5). Rectangular areas of interest (AOIs) were constructed around each food item in each trial and visual fixations were assigned to the corresponding item or coded as non-item fixations. If a non-item fixation was preceded and succeeded by fixations on the same item, the non-item fixation would also be assigned to this item. Other non-item fixations were not reassigned and discarded from all further analyses.

Krajbich 2011 This dataset is a straightforward extension of the Krajbich 2010 dataset to three-alternative choice. It contains data from 30 Caltech students, fulfilling the same inclusion criteria and receiving the same compensation as in Krajbich 2010.

Participants also performed the same liking rating task. The subsequent choice task again mirrors the Krajbich 2010 task in all aspects, except that now choices between triplets of food items were made. The items were arranged in a triangular fashion on the screen. In one half of the trials, this triangle pointed upwards (center option on top), in the other half it pointed downwards (center option at the bottom). Choices were indicated without time restriction and using the left, down and right arrow keys on a keyboard. Participants performed 100 choice trials. Eye movements were continuously recorded in the choice task, using a 50 Hz desktop-mounted Tobii eye tracker. The original preprocessing steps were identical to Krajbich 2010, except the cutoff criterion for missing fixation data was 500 ms, resulting in a total of 2966 remaining trials (mean $\pm$ s.e.m. number of trials dropped per participant was 1.1 $\pm$ 0.9).

Folke 2016 This dataset closely resembles the Krajbich 2011 three-alternative forced choice snack food task, but uses a gaze-contingent presentation format and a Becker-DeGroot-Marschak (BDM) procedure (Becker et al., 1964) to obtain willingness-to-pay (WTP) estimates for item value. The dataset contains data from 24 hungry participants. Participants were compensated with £15, less the price of one food item that they could buy during the experiment (see below). Participants first indicated their WTP between £0 and £3 for each of 72 common snack food items in a BDM procedure (Becker et al., 1964). In the following choice task, they made repeated choices without time limit between triplets of snack food items. Triplets were configured based on the participant's WTP estimates, so that an equal number of triplets contained three high-value items, three low-value items, two high-and one low-value, and two low- and one high-value item. This resulted in 48 different triplet configurations which were used once in each of three blocks, resulting in 144 trials per participant. Unlike in the Krajbich 2011 data, the task used a gaze-contingent stimulus presentation, where an item was only revealed while the participant's gaze was directed towards its position (indicated by a white-bordered box) on the screen. Also, items were arranged in a 2 x 2 grid with one random position left empty (not in a triangular shape as in Krajbich 2011). Participants gave confidence ratings on a visual analogue rating scale after each choice. However, we ignored the confidence ratings in our analyses, as they do not directly relate to gaze bias mechanisms. After the choice task, an auction based on the WTP estimates was held: For one of the chosen items, a "market price" was drawn from a uniform distribution between £0 and £3. If the participant's WTP for this item was above this drawn price, they would buy the item and receive a participation fee of £15 less the price paid. Otherwise, the item was not bought and participants received the full £15. Participants were asked to stay for another hour after the auction, and only allowed to eat the item they purchased (if any). Eye movements were continuously recorded during the choice task at 1000 Hz using an EyeLink 1000 Plus eye tracker. We obtained the data from the original author's repository in a preprocessed format. Original preprocessing steps included fixation detection and calculation of total fixation duration (dwell) towards each item.

Tavares 2017 This dataset is qualitatively different from the other three value-based choice datasets. Here, participants made perceptual judgments about the orientations

of two line segments, and their distance to a target. The dataset contains data from 25 participants, across four identical sessions on separate days. Participants received a $15 show-up fee each day, a $40 bonus for completing all sessions and an additional performance based bonus (see below). In contrast to the value-based choice datasets, there was no separate rating or bidding task. Within each session, there were 12 blocks of 28 trials. Each block began with a brief training session, where participants had to match single line segments to an exemplar and received feedback over their accuracy. Training terminated after six correct responses in a row. Then, a target line for the block was shown for 5 seconds. In the following choice trials, pairs of lines with orientations offset to the target were shown and participants had to choose the line closer to the target using the keyboard. There was no time limit. Target orientations were one of 20°, 35°, 55° or 70°. Stimulus lines were offset by −15°, −10°, −5°, 0°, 5°, 10°, or 15°. Stimulus pairs were constructed so that in each trial one line was closer to the target than the other. Each target orientation and configuration of stimulus pairs was presented equally often within each session. Each trial began with a 0.5 s forced fixation towards the center of the screen. A blue feedback box was shown around the chosen item for 1 s after a choice was made, but participants didn't receive feedback about the correctness of their choice. After every 5 trials, the block target was shown again. After each session participants received an additional $1 per correct choice among 25 randomly chosen trials. Eye movements were continuously recorded in the choice task, using a 500 Hz desktop-mounted EyeLink 1000 Plus eye tracker with head support. The data were obtained from the original authors in a preprocessed format. The original preprocessing steps included fixation detection and classifying them as fixations to the left or right item, or non-item fixations. As in Krajbich 2010 and Krajbich 2010 non-item fixations that were preceded and followed by fixations towards the same item were reclassified as fixations towards this item, and discarded otherwise. Trials with more than 50% of the RT recorded as non-item fixations were discarded. The mean percentage of discarded trials per participant was 5.2% (ranging from 0.15% to 25.8%).

Thomas 2020 This dataset is an extension of the Krajbich 2010 and Krajbich 2011 data to many-alternative choice. It contains data from 49 healthy English spearkers (17 female; 18-55 yrs, median: 23 yrs) who were equired to have normal or corrected-to-normal vision. Individuals wearing glasses or hard contact lenses were excluded from this study. Further, individuals were only allowed to participate, if they self-reportedly (I) fasted at least four hours prior to the experiment, (II) regularly ate the snack foods that were used in the experiment, (III) neither had any dietary restrictions nor (IV) a history of eating disorders and (V) didn't diet within the last six months prior to the experiment. Each participant completed the choice and liking rating task within a single session. In the choice task, participants were asked in each of 200 trials to choose the snack from item from sets of 9, 16, 25, or 36 items which they would like to eat most at the end of the experiment (50 trials per set size condition). Trials from the four set sizes were randomly intermixed. There were no time restrictions during the choice phase. Participant indicated when they had made a choice, by pressing the spacebar of a keyboard in front of them. A mouse cursor appeared in the center of the screen and participants had 3 s to click on their chosen item. Each choice set

was composed of 9, 16, 25 or 36 randomly selected snack food item images (random selection without replacement within a choice set). For each set size condition, these images were arranged in a square matrix shape, with the same number of images per row and column (3, 4, 5 or 6). Importantly, the average Euclidean distance between all item images in a set was the same across all four set size conditions (848 px). All images were displayed in the same size and resolution (164 x 133 px) and depicted a single snack food item centered in front of a consistent black background. The task used real snack food items that were familiar to the participants. During the choice task, participants' choices, RTs and eye-movements were recorded. Eye movement data were collected with a remote EyeLink 1000 system (SR Research Ltd., Mississauga, Ontario, Canada) with a sampling frequency of 500 Hz. Before the start of each trial, participants had to fixate a central fixation cross for at least 500 ms to ensure that they began each trial fixating on the same location. Fixation data were extracted from the output files obtained by the EyeLink software package (SR Research Ltd., Mississauga, Ontario, Canada) and classified classifying as either fixations to an item or non-item fixations. Am item gaze was defined as all consecutive fixations towards an item, without any fixation to other parts of the choice screen. We excluded complete trials from the analysis, if participants either chose an item that they didn't look at before pressing the spacebar, or if they clicked on the empty space between item images. These constraints were necessary to define the participant's choice as well as the time point of the choice (space bar press). The average number of trials dropped from the analysis was 3 (SE: 0.5) per participant and set size condition. In the subsequent liking rating task, participants indicated for each of the 80 snack foods that were used in the experiment, how much they would like to eat the item at the end of the experiment. Participants entered their ratings on a 7-point rating scale, ranging from -3 (not at all) to +3 (very much), with 0 denoting indifference. After the rating task, participants stayed for another 10 minutes and were asked to eat a single snack food item, which was selected randomly from one of their choices in the choice task. In addition to one snack food item, participants received a show-up fee of $10 and another $15 if they fully completed the experiment.

A.1.2 Erroneous response model

In line with existing drift diffusion model toolboxes (e.g., Wiecki et al., 2013), we include spurious trials at a fixed rate of 5% in all GLAM and hybrid model estimation procedures (see eq. A.2). We model these spurious trials with a subject-specific uniform likelihood distribution u_s. This likelihood describes the probability of a random choice for any of the N available items at a random time point in the range of a subject's empirically observed response times rt_s (cf. Ratcliff and Tuerlinckx (2002):

$$u_s(t) = \frac{1}{N(\max_{rt_s} - \min_{rt_s})} \tag{A.1}$$

$$l_{s,i}(t) = 0.95 \cdot p_i(t) + 0.05 \cdot u_s(t) \tag{A.2}$$

A.1.3 Item attributes

Liking rating: The liking rating of an item is defined by the rating that the subject assigned to this item in the respective liking rating task (see section 1.3.1 of the main text as well as Appendix A.1.1).

Position: This metric described the position of an item in a choice set and was encoded by two integer numbers: one indicating the row in which the item was located and the other indicating the respective column. Importantly, indices increased from left to right and top to bottom. For instance, in a choice set with 9 items, the column indices would be 1, 2, 3, increasing from left to right, while the row indices would also be 1, 2, 3, but increase from top to bottom. The item in the top left corner of a screen would therefore have a row and column index of 1, whereas the item in the top right corner would have a row index of 1 and a column index of 3.

Size: This metric describes the size of an item depiction with respect to the size of its image. In order to compute this statistic, we made use of the fact that all item images had the exact same absolute size and resolution within a dataset. First, we computed the fraction of the item image that was covered by the consistent black background. Subsequently, we subtracted this number from 1 to get a percentage estimate of how much image space is covered by the snack food item. As all item images had the same size and resolution, these percentage estimates are comparable across images.

A.1.4 Mixed effects modeling

All mixed-effects models were implemented and estimated using the bambi Python library (Yarkoni and Westfall, 2016). Bambi automatically generates weakly informative priors for all model terms by default (Westfall, 2017). We sampled two chains for each model, with at least 2000 samples each, using NUTS (Hoffman and Gelman, 2014). Convergence was diagnosed using the Gelman–Rubin criterion ($|\hat{R} - 1| < 0.05$) for all analyses. We declared fixed effects as statistically meaningful either when the 95% HDI excludes zero or when 95% of the posterior density is above (below) zero (see also ref. Cavanagh et al., 2014). In the latter case, we also report the proportion of the posterior mass above (below) zero, directly indicating the posterior probability of the effect being larger (smaller) than zero (see also ref. Kruschke, 2014).

Gaze influence on choice probability

Following previous work (Krajbich et al., 2010, Krajbich and Rangel, 2011), we first estimated a participant's probability of choosing an item in a choice set using logistic regression, based on its relative item value (the difference between the item's value and the mean value of all other items in that trial) and the range between the other items' values (this regressor was omitted in all two-item datasets). We then subtracted this estimated probability from the empirically observed choice (either 1 if the item was chosen or 0 otherwise) to obtain a residual choice probability that is corrected for the influence of the items' values.

To obtain an aggregate measure of the influence of gaze allocation on choice probability, we averaged the resulting residual choice probability for trials in which the item had a positive and negative final gaze advantage (computed as the difference in the fraction of the total fixation time that the participants spent fixating on the item and the average fraction that they spent fixating on the others). The difference between these two described the average difference in choice probability for the items with a positive versus negative final gaze advantage, when corrected for the influence of the values of the items.

A.1.5 Parameter estimation: GLAM

The GLAM was implemented in a Bayesian framework using the Python library PyMC3 (Salvatier et al., 2016). To reduce the influence of erroneous responses (for example, when the participant presses a button by accident or has a lapse of attention during the task) on parameter estimation, we explicitly included a model of contaminant processes in all estimation procedures (see Appendix A.1.2).

The GLAM has four parameters (v (ie., accumulation speed), γ (ie., gaze bias), σ (ie., accumulation noise), τ (ie., scaling of the relative evidence signals) as well as an optional parameter to extend the model for an additive influence of gaze on choice (ζ). We placed uninformative, uniform priors between sensible limits on all parameters as follows:

- $v \sim$ Uniform(1^{-10}, 0.01)
- $\gamma \sim$ Uniform(-10, 1)
- $\sigma \sim$ Uniform(1^{-10}, 0.02)
- $\tau \sim$ Uniform(0, 5)
- $\zeta \sim$ Uniform(0, 10)

The γ parameter has a natural upper bound at 1 (no multiplicative gaze bias) and the ζ parameter a natural lower bound at 0 (no additive gaze bias). The τ parameter has a natural lower bound at 0 (no sensitivity to differences in relative evidence). The v and σ parameters are also naturally bounded at 0.

All GLAM variants that were fitted to the Krajbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets were estimated with Markov Chain Monte Carlo sampling, by the use of the No-U-Turn-Sampler (NUTS; Hoffman and Gelman, 2014). We sampled two chains for each model with 500 tuning samples that were then discarded (burn-in), before drawing another 2000 additional posterior samples that we used to estimate the model parameters. If the sampler did not converge, as indicated by the Gelman–Rubin statistic ($|\hat{R} - 1| > 0.05$) or if the number of effective samples was low (< 100), the model was re-estimated using Metropolis sampling (two chains, with 10,000 samples each). Again, convergence was diagnosed using the Gelman–Rubin statistic and number of effective samples. Convergence was achieved for all models.

All GLAM varaints that were fitted to the Thomas 2020 dataset were estimated with Metropolis sampling, by sampling two chains with 5000 burn-in samples and another

5000 posterior samples that were used to estimate model parameters. Convergence was again checked by means of the Gelman–Rubin statistic ($|\hat{R}-1| < 0.05$) and the number of effective samples (> 100). If the sampler did not converge, we re-sampled the model and increased the number of burn-in samples by 5000 until convergence was achieved.

All reported parameter estimates are maximum a posteriori (MAP) estimates.

A.1.6 Parameter estimation: Probabilistic satisficing model

The probabilistic satisficing model was implemented in a Bayesian framework using the Python library PyMC3 (Salvatier et al., 2016). To reduce the influence of erroneous responses (for example, when the participant presses a button by accident or has a lapse of attention during the task) on parameter estimation, we explicitly included a model of contaminant processes in all estimation procedures (see Appendix A.1.2).

The probabilistic satisficing model model has five parameters which determine the influence of elapsed time (v) and cached value (α) onto its stopping probability, the strength of the additive (ζ) and multipicative (γ) gaze bias effecs on item value, and the sensitivity of its softmax choice rule (τ) (see section 1.3.4 of the main text). As for the GLAM (see Appendix A.1.5) we placed uninformative, uniform priors between sensible limits on all parameters as follows:

- $v \sim$ Uniform(0, 0.001)
- $\alpha \sim$ Uniform(0, 0.001)
- $\gamma \sim$ Uniform(0, 1)
- $\zeta \sim$ Uniform(0, 10)
- $\tau \sim$ Uniform(0, 10)

The γ parameter has a natural upper bound at 1 (no multiplicative gaze bias) and the ζ parameter a natural lower bound at 0 (no additive gaze bias). The τ parameter has a natural lower bound at 0 (no sensitivity to differences in relative evidence), while the v and α parameters are also naturally bounded at 0. To avoid any negative values of the stopping probability, we further re-scaled all item values of the Thomas 2020 dataset to a range from 1-7 and bounded γ at 0.

We fitted the probabilistic satisficing model by the use of Markov Chain Monte Carlo Metropolis sampling and sampled two chains with 5000 burn-in samples that were discarded before another 5000 posterior samples were drawn that were used to estimate model parameters. Convergence of the the parameter traces was checked by means of the Gelman–Rubin statistic ($|\hat{R} - 1| < 0.05$) and the number of effective samples (> 100). If the sampler did not converge, we re-sampled the model and increased the number of burn-in samples by 5000 until convergence was achieved.

All reported parameter estimates are maximum a posteriori (MAP) estimates.

A.1.7 Parameter estimation: Independent evidence accumulation model

The independent evidence accumulation model was implemented in a Bayesian framework using the Python library PyMC3 (Salvatier et al., 2016). To reduce the influence of erroneous responses (for example, when the participant presses a button by accident or has a lapse of attention during the task) on parameter estimation, we explicitly included a model of contaminant processes in all estimation procedures (see Appendix A.1.2).

The independent evidence accumulation model has four parameters (v (ie., accumulation speed), γ (ie., multiplicative gaze bias), ζ (ie., additive gaze bias), σ (ie., accumulation noise); see section 1.3.4 of the main text). We placed uninformative, uniform priors between sensible limits on all parameters as follows:

- $v \sim \text{Uniform}(1^{-7}, 0.005)$
- $\gamma \sim \text{Uniform}(0, 1)$
- $\sigma \sim \text{Uniform}(1^{-7}, 0.05)$
- $\zeta \sim \text{Uniform}(0, 10)$

The γ parameter has a natural upper bound at 1 (no multiplicative gaze bias) and the ζ parameter a natural lower bound at 0 (no additive gaze bias). The v and σ parameters are also naturally bounded at 0. To avoid any negative drift terms in the race we further bounded γ at 0.

We fitted the independent evidence accumulation model by the use of Markov Chain Monte Carlo Metropolis sampling and sampled two chains with 5000 burn-in samples that were discarded before another 5000 posterior samples were drawn that were used to estimate model parameters. Convergence of the the parameter traces was checked by means of the Gelman–Rubin statistic ($|\hat{R} - 1| < 0.05$) and the number of effective samples (> 100). If the sampler did not converge, we re-sampled the model and increased the number of burn-in samples by 5000 until convergence was achieved.

All reported parameter estimates are maximum a posteriori (MAP) estimates.

A.1.8 Model simulations

Choice and RT data were simulated from each model according to the following procedures. Each trial in the dataset for which simulations are to be obtained was repeated 50 times. For every trial, the model used the observed item values and gaze distributions. With a fixed rate of 5% the simulation produced a random choice and RT between the participant's minimum and maximum observed RT (see eq. A.1). With a rate of 95% the choice and RT were simulated from the respective model. We defined the model parameter estimates that were used for the simulation as the maximum a posteriori estimates (MAP) of the posterior traces of the individual models (see Appendix A.1.5, A.1.6, and A.1.7).

A.2 Figures

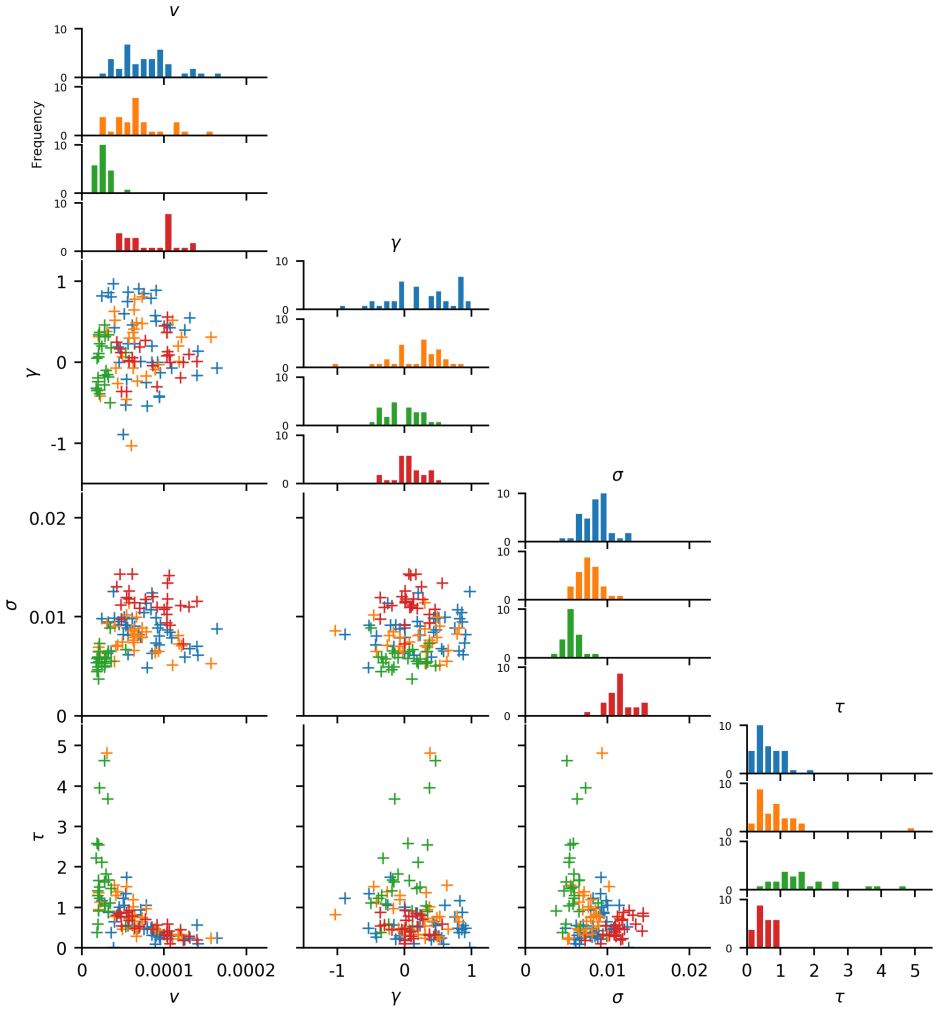

Figure A.1: Parameter estimates from the in-sample GLAM fits and their relationships. See Appendix Table A.2 for an overview of parameter correlations. Colours indicate datasets (blue: Krajbich 2010, orange: Krajbich 2011, green: Folke 2016, red: Tavares 2017)

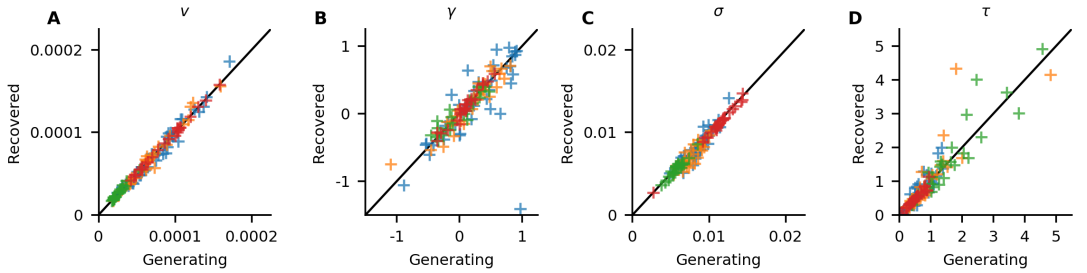

Figure A.2: GLAM model recovery. We performed a parameter recovery study to validate the parameter estimates. We simulated data using the GLAM and the corresponding individual parameter estimates for the Krajbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets (see section 1.4.1 of the main text and Appendix Fig. A.1). For each empirically observed trial (i.e., set of item values and gazes) the GLAM simulated choice and RT data once, resulting in a GLAM-generated dataset that matches the original data in size andstructure. We then performed the exact same parameter estimation procedure as we did in our main analyses. Ideally, the recovered parameters should match the originally estimated (data-generating) ones. We performed a bias analysis similar to the bias analysis of observed and predicted data in the main text (see Fig. 1.5 of the main text): Using a mixed effects model with random intercepts and slopes for each dataset, we regressed the parameter estimates on an indicator variable, indicating whether a parameter estimate was generating data or estimated from data. All parameters could be recovered to a satisfying degree without bias (A: $\beta = -4e^{-09}$, 95% HDI = [-1e-05, 1e-05] difference between the generating and recovered v values; B: $\beta = -0.03$, 95% HDI = [-0.19, 0.12] difference between the generating and recovered γ values; C; $\beta = -3e^{-04}$, 95% HDI = [$-1e^{-03}$, $4e^{-4}$] difference between the generating and recovered σ values; D: $\beta = 0.04$, 95% HDI = [-0.25, 0.34] difference between the generating and recovered τ values). Colours indicate datasets (blue: Krajbich 2010, orange: Krajbich 2011, green: Folke 2016, red: Tavares 2017)

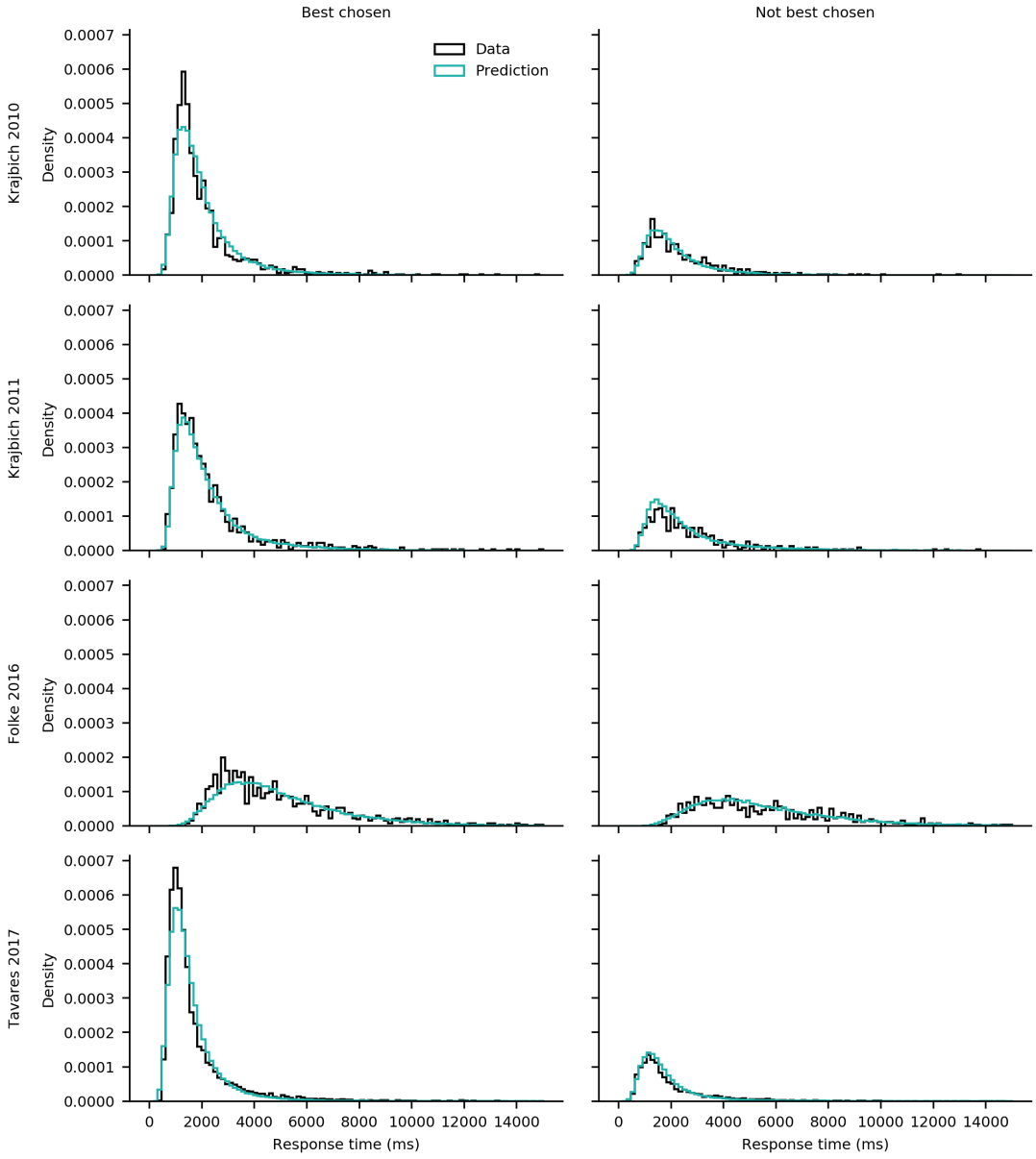

Figure A.3: Observed and out-of-sample predicted aggregate RT distributions for all odd-numbered trials for the Krajbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets. Each row shows data and predictions of one dataset. The left column shows RTs for trials where the highest rated item was chosen (n= 16182), the right column shows RTs for trials where any other item was chosen (n= 4831).

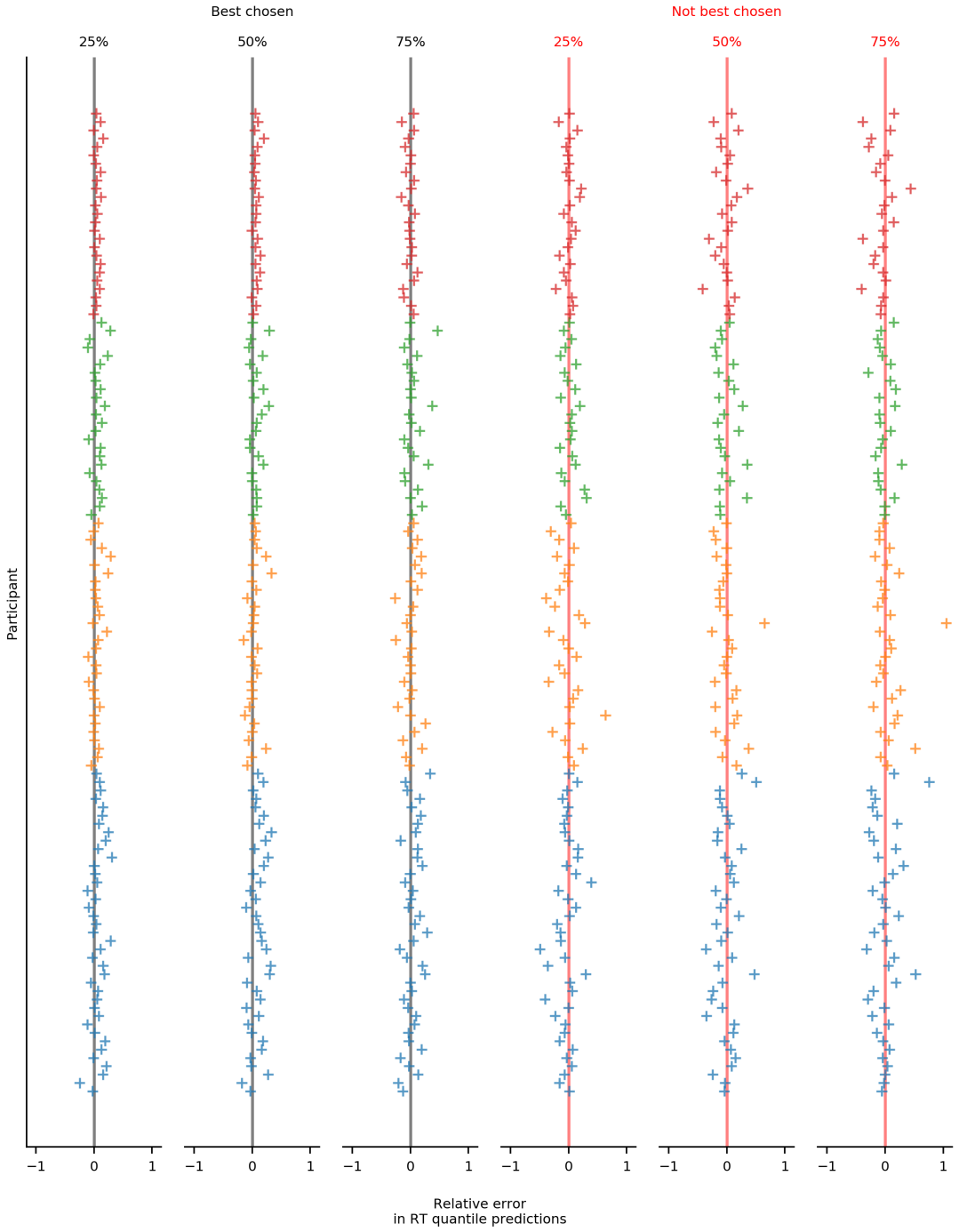

Figure A.4: Individual observed and out-of-sample predicted RT distributions for all odd-numbered trials of the Krajbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets. Plotted are relative errors (difference between predicted and observed, relative to observed values) in predicted vs. observed RT quantiles (25%, 50% and 75%) for each individual (n= 118). The left three columns show RT quantiles for trials where the highest rated item was chosen (n= 16182), the right three columns show RT quantiles for trials where any other item was chosen (n= 4831). Colours indicate datasets (blue: Krajbich 2010, orange: Krajbich 2011, green: Folke 2016, red: Tavares 2017).

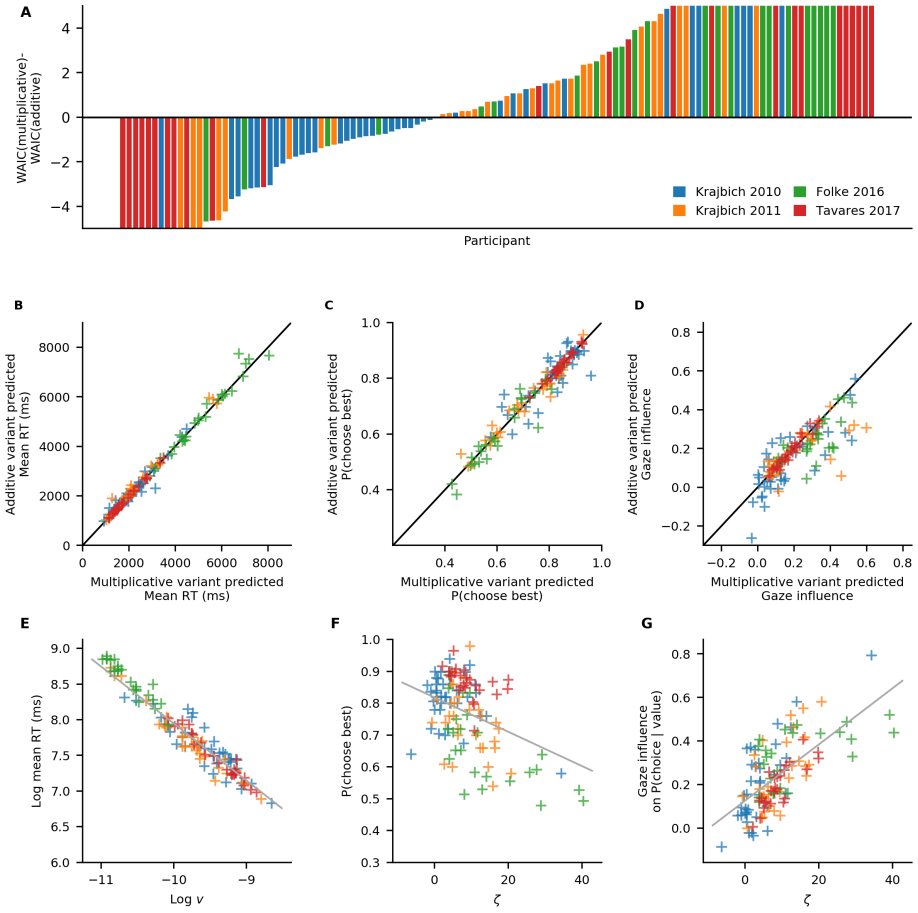

Figure A.5: Comparison of the individual response predictions of two GLAM variants with multiplicative and additive gaze bias mechanisms in the Karjbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets. Both model variants fit comparable portions of the participants better in terms of WAIC (A), and make highly similar out-of-sample predictions on all behavioural metrics (B-D). A: Individual differences in widely applicable information criterion (WAIC) between the additive and multiplicative variant. Negative differences indicate better fits of the multiplicative variant. B-D: Comparison of the individual out-of-sample response predictions of the multiplicative and additive GLAM variants for all odd-numbered trials. Both model's predictions mimic each other considerably in mean RT (A; β = 54 ms, 95% HDI = [-321 ms , 446 ms] difference in predicted mean RT), probability of choosing the best item (B; β = -1.11%, 95% HDI = [-4.01, 3.96] difference in predicted probability of choosing the best item) and influence of gaze on choice probability (C; β = -3.65%, 95% HDI = [-9.54, 1.56] difference in predicted gaze influence on choice probability). Both model variants were fitted to the even-numbered empirical trials. Further, the parameter estimates of the additive GLAM variant show a similar association with individuals' response behaviour as the parameter estimates of the multiplicative variant (Fig. 1.6 of the main text): E: Log-transformed mean RTs decrease with increasing log-transformed v-estimates (β = -0.79, 95% HDI = [-0.87, -0.73]). D: Individuals' probability of choosing the best item decreases with increasing ζ estimates (β = 0.54%, 95% HDI = [-0.97, -0.06]). G: Behavioural gaze influence measures increase with increasing ζ estimates (β = 1.29%, 95% HDI = [0.27, 2.21]). Note that the y-axis in A is truncated to better show distribution around 0. WAIC differences range from -35.91 to 56.60.

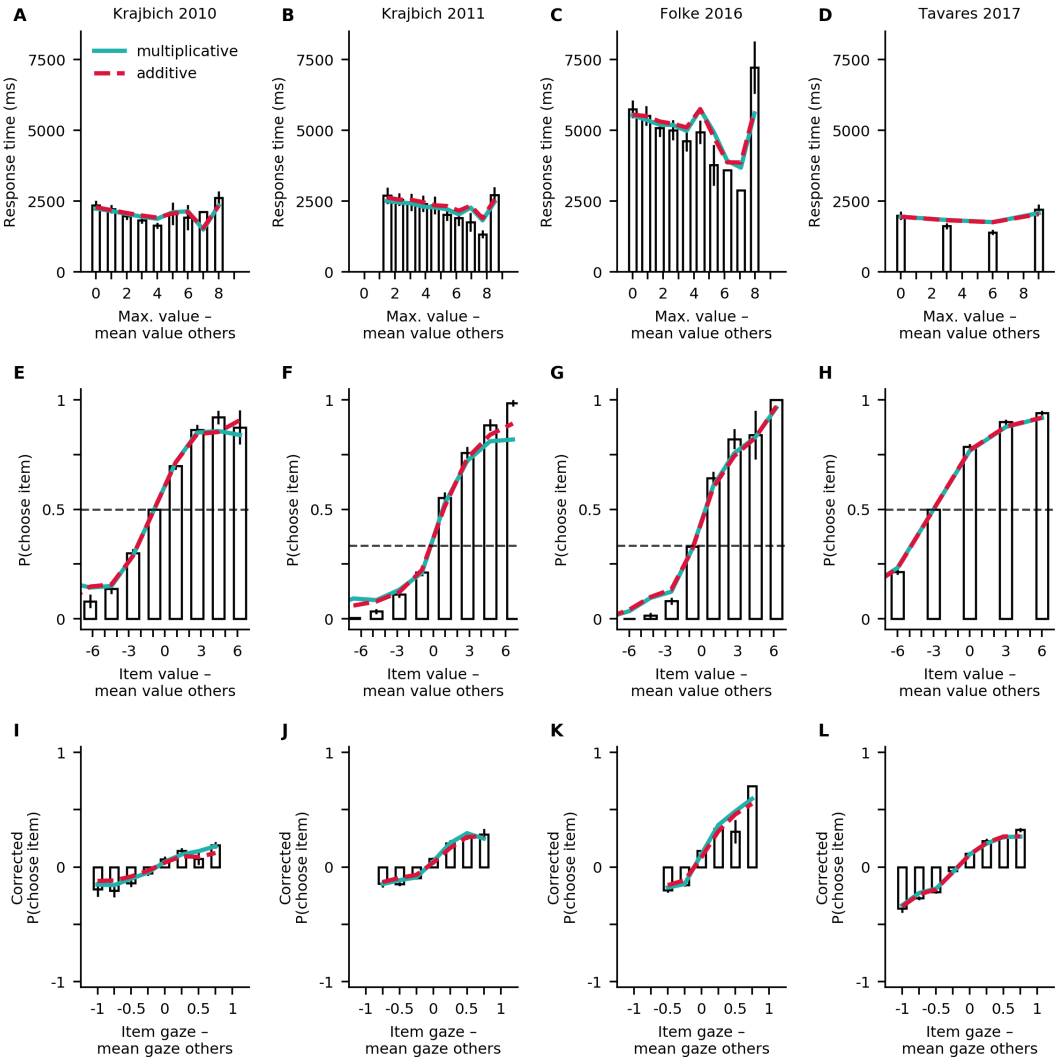

Figure A.6: Comparison of group-averaged out-of-sample predictions of two GLAM variants with multiplicative and additive gaze bias mechanisms for all odd-numbered trials of the Karjbich 2010, Krajbich 2011, Folke 2016, and Tavares 2017 datasets. Both model variants capture the data similarly well, across different behavioural metrics and datasets. A-D: Mean RTs as a function of the difference between the maximum itemvalue of a choice set and the mean value of all other items. E-H: Mean probability of choosing an item as a function of the difference between the item's value and the mean value of all other items. Dashed horizontal lines indicate random choice probabilities. I-L: Mean probability of choosing an item as a function of the difference between the gaze towards this item and the mean gaze to all other items, when corrected for the influence of item value on choice probability (see Appendix A.1). Both model variants were fitted to the even-numbered empirical trials. Colours indicate model variants, with solid green lines representing the model variant with a multiplicative effect of gaze on choice behaviour and dashed redlines representing the model variant with an additive effect of gaze on choice. Observed data is shown as bars. Vertical lines indicate standard errors.

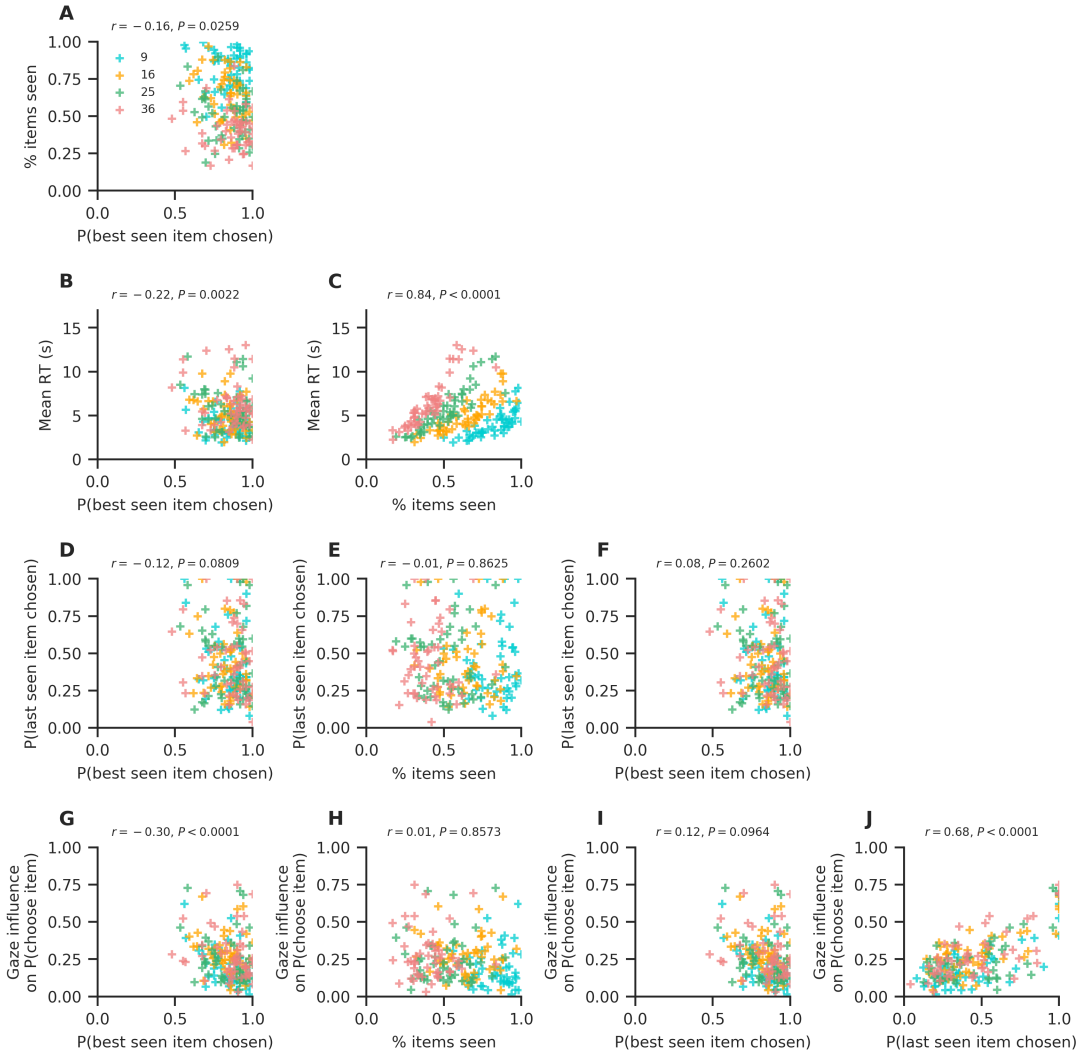

Figure A.7: Associations of the MAFC choice psychometrics presented in Fig. 1.7 F of the main text. Crosses indicate subject means, while colors indicate choice set sizes. Pearson's r correlation with corresponing P-value is indicated for each association.

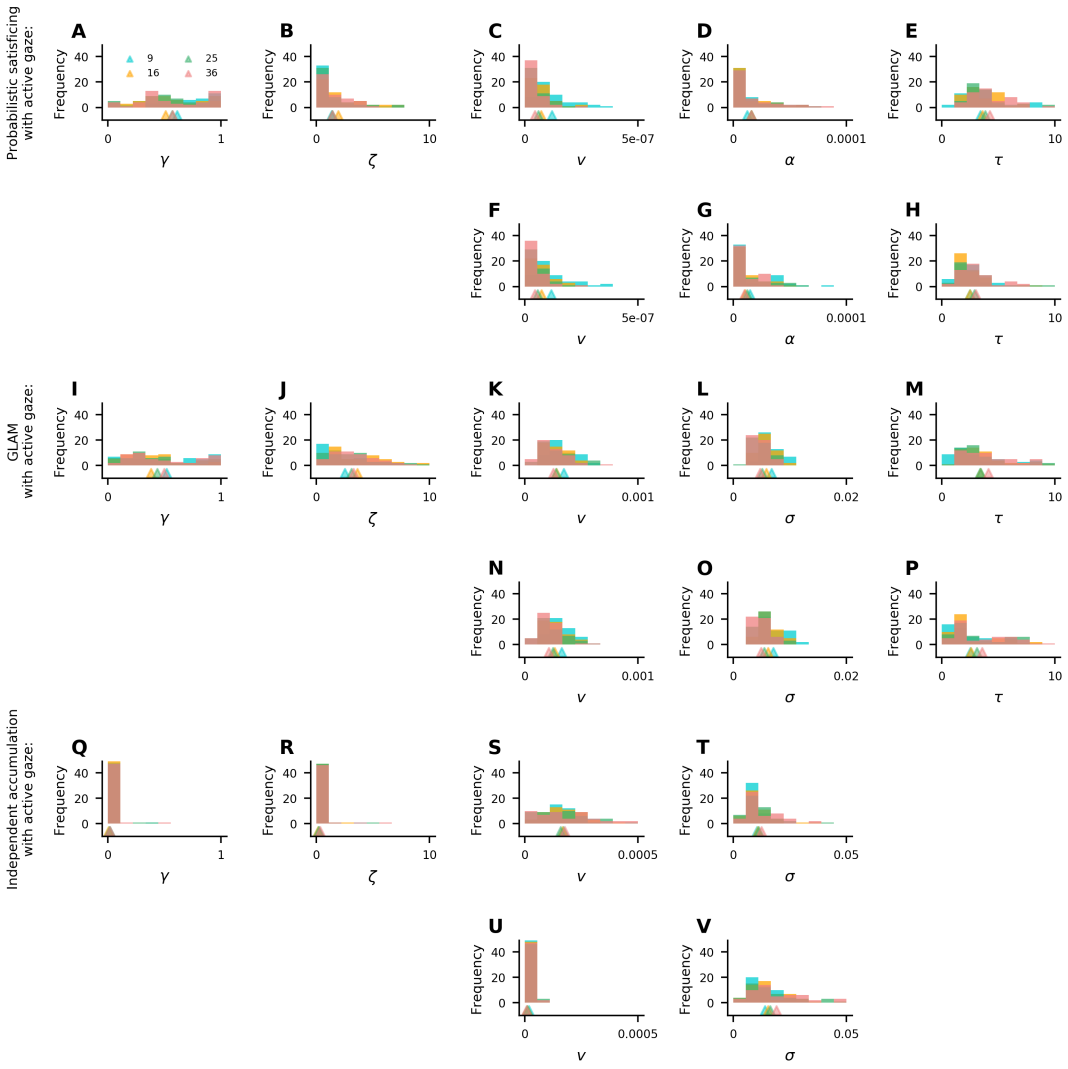

Figure A.8: Parameter estimates from the in-sample fits of the probabilistic satisficing model (active gaze variant: A-E, passive gaze variant: F-H), GLAM (active gaze variant: I-M, passive gaze variant: N-P), and independent evidence accumulation model (active gaze variant: Q-T, passive gaze variant: U-V) for the Thomas 2020 dataset. Colors indicate choice set size conditions (9: blue, 16: yellow, 25: green, 36: red). Triangles indicate the mean parameter estimate in each set size condition.

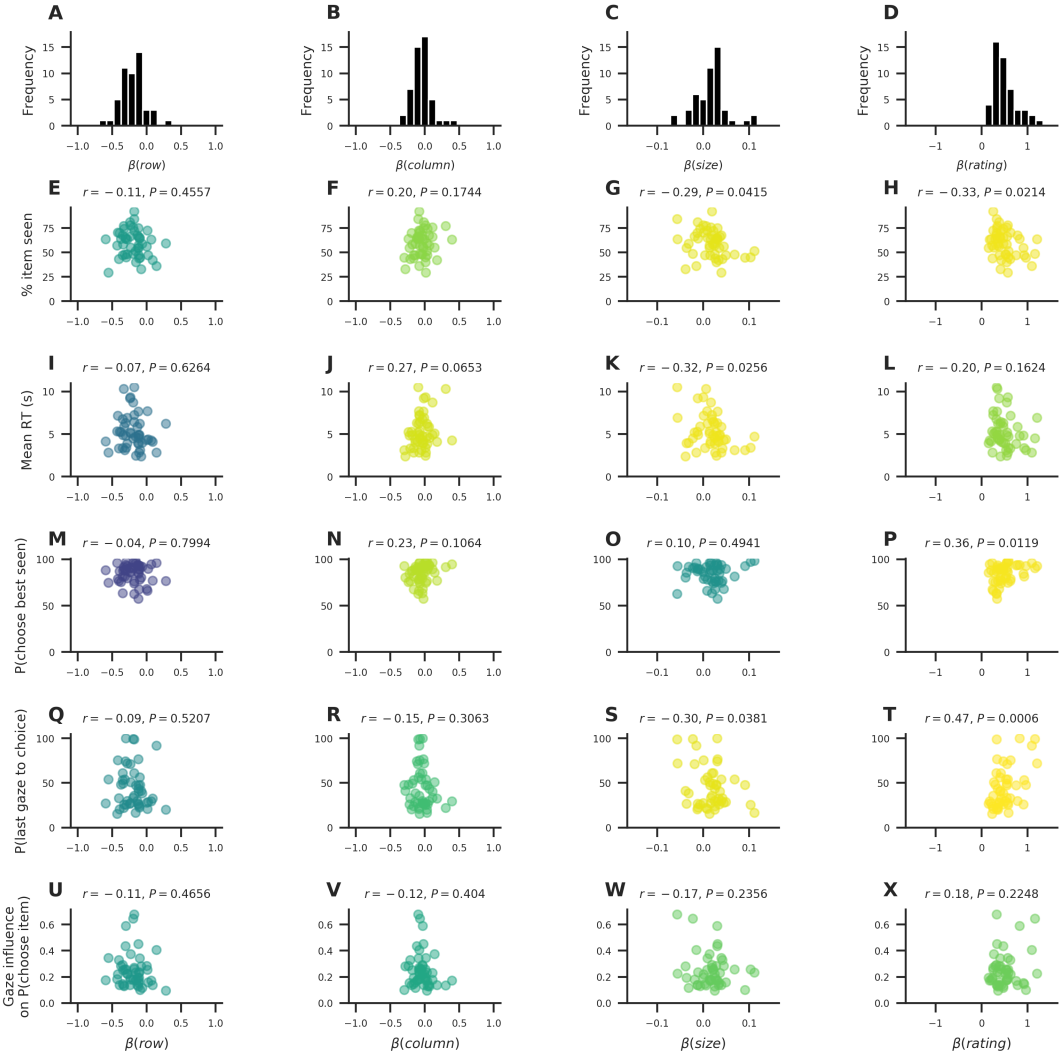

Figure A.9: Association between the MAFC choice psychometrics presented in Fig. 1.7 of the main text and a set of measures describing individuals' visual search behaviour. To quantify individuals' visual search, we computed a mixed effects regression model for each individual in the data, estimating how much the individual's allocation of gaze to an item (measured as the fraction of trial time that the item was looked at) is influenced by the item's attributes (namely, the item's row- and column-position, size, and liking rating; see Appendix A.1.3) as well as the choice set size. We then studied the relationship between the resulting regression β estimates (A-D) for each of the item attributes and each individuals' overall mean on the five behavioral choice metrics of presented in Fig. 1.7 of the main text (namely, the mean fraction of items looked at in a trial, mean RT, the probability of choosing the best seen item from a choice set, the probability of looking at the chosen item last, and the gaze influence measure). Pearson's r correlation is indicated for each association. Brighter yellow colors indicate smaller P-values. Scatter points indicate subjects.

A.3 Tables

Dataset	Krajbich '10	Krajbich '11	Folke '16	Tavares '17	Overall
N	39	30	24	25	118
Set size	2	3	3	2	-
Choice domain	value-based	value-based	value-based	perceptual	-
Mean RT (ms)	2,192 ms (851 ms)	2,462 ms (1298 ms)	5,414 ms (1,284 ms)	1,849 ms (601 ms)	2,844 ms (1,676 ms)
P(choose best)	81% (6%)	72% (10%)	66% (12%)	86% (5%)	77% (12%)
Gaze influence	19% (17%)	25% (14%)	35% (11%)	19% (9%)	24% (15%)

Table A.1: Means are given, with respective standard deviations in parentheses.

.

Parameter	v	γ	τ	σ
v	1			
γ	0.06 ($P = 0.50$)	1		
τ	-0.59 ($P < 0.001$)	-0.10 (P = 0.29)	1	
σ	0.26 ($P = 0.004$)	0.10 ($P = 0.29$)	0.27 ($P = 0.003$)	1

Table A.2: Pearson correlation coefficients ($n = 118$) for the GLAM's in-sample parameter estimates. Reported P-values come from two-tailed tests.

Appendix: The analysis of fMRI data through deep learning models

B.1 Methodology

B.1.1 HCP experiment task details

Working Memory Each of the two runs of the working memory task consisted of eight task (25 s each) and four fixation blocks (15 s each). In each of the task blocks, particpants saw images of one of four different stimulus types (namely, images of body parts, faces, places or tools). These four stimulus types are known to reliably engage distinct cortical regions (Downing et al., 2001) across subjects Peelen and Downing (2005) and time (Fox et al., 2009). Half of the task blocks used a 2-back working memory task (participants were asked to respond "target" when the current stimulus was the same as the stimulus 2 back) and the other half a 0-back working memory task (a target stimulus was presented at the beginning of each block and participants were asked to respond "target" whenever the target stimulus was presented in the block). Each task block consisted of 10 trials (2.5 s each). In each trial, a stimulus was presented for 2 s followed by a 500 ms interstimulus interval (ISI). We were not interested in identifying any effect of the N-back task condition on the evoked brain activity and therefore pooled the data of both N-back conditions.

Gambling Participants played a card guessing game in which they were asked to guess the number on a mistery card. The potential card numbers ranged from 1 to 9 and participants were asked to indicate whether they think that the number is going to be above or below 5. Participants received feedback in form of a number on the card. Importantly, the number on the card was dependent on whether the respective trial was marked as a reward, loss, or neutral trial. In addition to the number, the feedback included a green arrow pointing upwards with "1" for reward trials or a red arrow pointing downwards next to "−0.50" for loss trials or the number "5" and a gray double headed arrow for neutral trials. Participants had 1.5 s to indicate a guess (during this time a "?" was presented), while the subsequent feedback was presented for 1.0 s. In addition, there was a 1.0 s intertrial interval with a "+" on the screen. The task was presented in blocks that each included eight trials that were either mostly reward (6 rewards trials that were pseudo-randomly interleaved with either 1 neutral and 1

loss trial, 2 neutral trials, or 2 loss trials) or mostly loss (6 loss trials interleaved with either 1 neutral and 1 reward trial, 2 neutral trials, or 2 reward trials) trials. In each of the two fMRI runs there were 2 mostly reward and 2 mostly loss blocks, interleaved with 4 fixation blocks (15 s each, during which a "+" is presented on the screen). All participants were provided with money as a result of completing the experiment. The amount they received was standardized due to the fixed nature of the experiment.

Motor Participants were presented with visual cues that asked them to tap their left or right fingers, squeeze their left or right toes, or move their tongue. The task was presented in blocks of 12 s that each included only one movement type (10 movements). Each block was preceded by a 3 s cue. In each of the two fMRI runs, 13 blocks were presented with 2 blocks for tongue movements, 4 blocks for hand movements (2 left, 2 right), and 4 blocks for foot movements (again, 2 left and 2 right). In addition, three 15 s fixation blocks were included in each run.

Language This task consisted of two runs that each interleaved 4 blocks of a story task and 4 blocks of a math task. In the story task, participants were presented with brief auditory stories (5-9 sentences) that were adapted from Aesop's fables. After each story, a 2-alternative forced-choice question asked the particpant about the topic of the story. In the math task, particpants were similarly presented with an auditory math problem that asked them to complete 2-alternative forced choice addition or subtraction problems. For example, participants heard the operation "fourteen plus twelve", followed by "equals" and then two choice alternatives ("twenty-nine or twenty-six"). Participants indicated with a button press whether they choose the first or second answer. The lengths of the blocks varied (with an average of approximately 30 s per block), but the task was designed in such a way that the math task blocks matched the length of the story task blocks (with some additional math trials at the end of a block if needed to complete the 3.8min run).

Social Participants were presented with video clips (20 s each) that showed objects (squares, circles, triangles) that either interacted in some way or were moving randomly. After each video clip, participants indicated whether they think that the objects had a social interaction (an interaction that appears as if the objects are taking into account each other's feelings or thoughts), they are not sure, or they think the objects did not interact. Each of the two fMRI runs included 5 video blocks (2 with interaction and 3 without in one run and 3 with interaction and 2 without in the other run) as well as 5 15 s fixation blocks.

Relational In this task, participants saw stimuli that were composed of six different shapes that were filled with one of six different textures. In the relational task condition, 2 pairs of objects were presented, one at the top of the screen and the other at the bottom. Participants were told that they should first decide what dimension (shape or texture) differs across the top pair of objects and then whether the bottom pair of objects differs along the same dimension. In the matching condition, participants were

shown two objects at the top of the screen and one at the bottom. A word in the middle of the screen then indicated whether participants should decide if the bottom object matched either of the two top objects on the "shape" or "texture" dimension. In the relational condition, stimuli were presented for 3500 ms, with a 500 ms intertrial interval and four trials per block. In the matching condition, stimuli were presented for 2800 ms, with a 400 ms intertrial interval, and a total of five trials per block. Each block lasted a total of 18 s. In each of the two fMRI runs three relational blocks, three matching blocks and three fixation blocks (16 s each) were presented.

Emotion Participants were presented with with two faces at the bottom of the screen and one face at the top. These faces had an either angry or fearful expression. The participants were asked to decide which of the two faces on the bottom matches the face at the top. In the control trials, participants were asked to decide which of two shapes at the bottom of the screen matches a shape that is presented at the top. In this task, trials were presented in blocks of six trials of the same task (face or shape). In each trial, the stimulus was presented for 2 s in addition to a 1 s intertrial interval. Each block was further preceded by a 3 s cue for the task (shape or face). Each of the two fMRI runs included three face and three shape blocks. Due to a bug in the experiment script, the experiment stopped before the final three trials of the last block of each trial (for further details on this bug, see Barch et al. (2013)).

B.1.2 DeepLight architectures of study I

In study I, we specified the sequence of 2D-convolution layers of 2D-DeepLight as follows: conv3-16(1), conv3-16(1), conv3-16(2), conv3-16(1), conv3-32(2), conv3-32(1), conv3-32(2), conv3-32 [notation: conv(kernel size) - (number of kernels)(stride size)]. We further applied zero-parding to the outputs of each convolution layer, to avoid any loss of dimensionality between the convolution layers when the stride size is not larger than one voxel. Thereby, adding zeros to the borders of the inputs to each convolution layer so that the outputs of the convolution layers have the same dimensionality as their inputs, if a stride of one voxel is applied, and only decrease in size, when a larger stride is used. Each of the two subsequent LSTM units contained 40 output neurons. To make a decoding decision, both LSTM units passed their output for the last sequence element to a fully-connected softmax output layer with one neuron for each of the cognitive states in the data.

B.1.3 DeepLight architectures of study II

2D-DeepLight In study II, we specified the sequence of 2D-convolution layers of the 2D-DeepLight architecture as follows: conv3-8(1), conv3-8(1), conv3-16(2), conv3-16(1), conv3-32(2), conv3-32(1), conv3-32(2), conv3-32(1), conv3-64(2), conv3-64(1), conv3-64(2), conv3-64(1) (notation: conv(kernel size) - (number of kernels)(stride size)). We further applied zero-parding to the outputs of each convolution layer, to avoid any loss of dimensionality between the convolution layers when the stride size is not larger than one voxel. Thereby, adding zeros to the borders of the inputs to each convolution layer so that the outputs of the convolution layers have the same dimensionality as their

inputs, if a stride of one voxel is applied, and only decrease in size, when a larger stride is used. The two subsequent LSTM units contained 64 output neurons respectively, while the fully-connected softmax output layer contained one neuron for each cognitive state in the data.

3D-DeepLight In study II, we specified the sequence of 3D-convolution layers of the 3D-DeepLight architecture as follows: conv3-8(1), conv3-8(1), conv3-8(2), conv3-8(1), conv3-16(2), conv3-16(1), conv3-32(2), conv3-32(1), conv3-64(2), conv3-64(1), conv3-128(2), conv3-128(1) (notation: conv(kernel size) - (number of kernels)(stride size)). We further applied zero-parding to the outputs of each convolution layer, to avoid any loss of dimensionality between the convolution layers when the stride size is not larger than one voxel. Thereby, adding zeros to the borders of the inputs to each convolution layer so that the outputs of the convolution layers have the same dimensionality as their inputs, if a stride of one voxel is applied, and only decrease in size, when a larger stride is used. To make a decoding decision, 3D-DeepLight passes the higher-level representation of brain activity resulting from the 3D-convolutional feature extractor to a one-dimensional convolution layer with one kernel for each cognitive state in the data. Lastly, a pooling layer averages over the resulting values within each kernel-dimension (resulting in one value for each target cognitive state) and passes these through a softmax output function.

B.1.4 Outer brain mask used in study I

In line with previous work (Jang et al., 2017), we applied an outer brain mask to each fMRI volume. We first identified those voxels whose activity was larger than 5% of the maximum voxel signal within the fMRI volume and then only kept those voxels for further analysis that were positioned between the first and last voxel to fulfill this property in the three spatial dimensions of any functional brain volume of our dataset. This resulted in a brain mask spanning 74 $\times$ 92 $\times$ 81 voxels (X $\times$ Y $\times$ Z).

B.1.5 Parameter estimation: Conventional analysis approaches

Throughout the following, we will describe the procedures that we used to estimate the parameters (and thereby a set of brain maps for each cognitive state) of each of the three baseline analysis approaches used in study I.

The data of many subjects (with approximately 1GB per subject) can easily exceed the working memory capacities of a regular working station. For this reason, we adapted the parameter estimation procedures for the searchlight analysis and whole-brain lasso, when switching from the subject- to the group-level.

General Linear Model: Our GLM analyses included one predictor for each of the four cognitive states in the design matrix (each representing a box-car function for the occurrence of a cognitive state; for methodological details on the GLM, see section 2.3.2 of the main text). We convolved these predictors with a canonical haemodynamic

response function (Lindquist et al., 2009, HRF;), as implemented in NiPy 0.4.1 (Gorgolewski et al., 2011), to generate the model predictors. We added temporal derivative terms derived from each predictor, an intercept and an indicator of the experiment run to the design matrix, which we all treated as confounds of no interest. The derivative terms were computed by the use of the cosine drift model as implemented in NiPy 0.4.1 Gorgolewski et al. (2011). All β-coefficients and error terms of the GLM analysis were estimated by the use of a first-level autoregressive model, as implemented in NiPy 0.4.1 Gorgolewski et al. (2011).

To generate a set of subject-level brain maps, we computed a linear first-level contrast within the data of each individual subject (representing a linear contrast between one of the cognitive states and all others). The resulting brain maps indicate the estimated Z-values of these contrasts.

To generate a set of group-level brain maps with the GLM, we computed a second-level GLM contrast by the use of the standard two-stage procedure for a random-effects group-level analysis, as proposed by Holmes and Friston (1998). Here, the subject-level regression coefficients β are treated as random effects in a second-level linear contrast analysis, where the distribution of first-level β-contrasts is assessed by the use of a one-sample t-test (again, contrasts were computed between each cognitive state and all others). The resulting group-level brain maps show the t-values resulting from this test.

Searchlight Analysis: On the subject-level, we trained the searchlight analysis in a one-vs-rest procedure. Here, one SVM classifier is trained at each location in the brain to distinguish each cognitive state from all others. A decoding decision is then made according to the classifier with the most certainty that the sample belongs to its respective cognitive state. We first trained the searchlight analysis within the data of the first experiment run of a subject and subsequently predicted the cognitive states underlying the data of the second experiment run. The resulting brain maps indicate the decoding accuracies achieved by each of these SVM classifiers in the second experiment run at each searchlight location in the brain.

For the group-level searchlight analysis, we trained and evaluated the searchlight on the β-coefficient maps of a first-level GLM analysis of each individual subject (resulting in one β-coefficient map per subject and cognitive state). This is a common approach for group-level predictions with the searchlight analysis and is widely applied in the neuroscience literature (e.g., Helfinstein et al., 2014, Reverberi et al., 2018, Schuck et al., 2016). First, we trained the searchlight analysis in a one-vs-rest procedure on the subject-level β-coefficient maps of the training dataset. Subsequently, we used each of the trained searchlight classifiers to decode the cognitive states underlying each subject-level β-coefficient map in the test data. The resulting group-level brain maps represent the decoding accuracies achieved by each of these searchlight classifiers in the test data at each location in the brain.

Whole-brain Least Absolute Shrinkage Logistic Regression: On the subject-level, we trained the whole-brain lasso in a one-vs-rest procedure. To determine the magnitude of the regularization parameter λ, we additionally applied a grid search: First, we split the full training data of a subject, containing the data of the first exper-

iment run, into the eight experiment blocks of this run (two per cognitive state). We then separated these blocks into a new training dataset (containing the first experiment block of each cognitive state) and a new validation dataset (containing the second experiment block of each cognitive state). Subsequently, for each λ-value of the parameter grid, we fitted the whole-brain lasso to the data of the newly formed training dataset and evaluated its performance on the new validation data. Importantly, we utilized a logistic model implementation of the scikit-learn python library (Abraham et al., 2014). Here, the regularization parameter (C) is implemented inversely to the regularization strength λ (with lower values indicating stronger regularization; see eq. 2.3 of the main text). With this procedure, we evaluated a grid of 100 logarithmically spaced C-values between $1e-6$ and 100. From these values we then selected the C-parameter for the subject that achieved the highest decoding accuracy in the new validation dataset (for an overview of the selected subject C-parameters, see Appendix Table B.1). Subsequently, we used the selected C-value to fit the whole-brain lasso to the full training data of the subject (containing the entire data of the first experiment run). The resulting brain maps of the whole-brain lasso show the coefficient estimates of each of these one-vs-rest logistic models.

On the group-level, we trained the whole-brain lasso in a stochastic gradient descent learning procedure (Kiefer and Wolfowitz, 1952). Here, the regularized logistic model is fit iteratively to subsets of the full training data. At each iteration, the gradient of the loss function is estimated and the model's parameters are updated accordingly. To determine the strength of the regularization parameter λ, we again applied a grid search procedure. For each value of the λ-grid, we trained the whole-brain lasso over 25 epochs. In each epoch, we randomly selected the fMRI data of five subjects from the training dataset (see section 2.4.1 of the main text). We then randomly drew 50 batches, each containing 50 randomly drawn fMRI volumes, and updated the whole-brain lasso parameters iteratively for each batch. After completing the 25 epochs, we evaluated the decoding performance of the whole-brain lasso on the full test dataset. Overall, we evaluated 20 different λ-parameters in this grid-search and selected the λ-value achieving the highest decoding accuracy in the test dataset (λ=0.0001; for an overview of the evaluated λ-values and resulting decoding accuracies, see Appendix Table B.2). We then used the selected λ-parameter to train the whole-brain lasso in the same stochastic gradient procedure described before. This time, however, spanning 200 training epochs. The group-level brain maps of the whole-brain lasso represent the resulting one-vs-rest logistic model coefficients.

B.1.6 Parameter estimation: DeepLight, study I

We iteratively trained DeepLight on the training dataset (for details on the training data of study I, see section 2.4.1 of the main text) through backpropagation (Hecht-Nielsen, 1992) over 60 epochs by the use of the ADAM optimization algorithm as implemented in tensorflow 1.4 (Abadi et al., 2016). To prevent overfitting, we applied dropout regularization to all network layers (Srivastava et al., 2014), global gradient norm clipping (with a clipping threshold of 5; Pascanu et al., 2013), as well as an early stopping of the training (for an overview of training statistics, see Appendix Fig. B.1).

During the training, we set the dropout probability to 50% for all network layers, except for the first four convolution layers, where we reduced the dropout probability to 30% for the first two layers and 40% for the third and fourth layer. Each training epoch was defined as a complete iteration over all samples in the training dataset. We used a learning rate of 0.0001 and a batch size of 32. All network weights were initialized by the use of a normal-distributed random initialization scheme (Glorot and Bengio, 2010). The DL model was written in tensorflow 1.4 (Abadi et al., 2016) and the interprettensor library (https://github.com/VigneshSrinivasan10/interprettensor).

B.1.7 Parameter estimation: DeepLight, study II

2D-DeepLight We iteratively trained the 2D-DeepLight architecture through back-propagation (Hecht-Nielsen, 1992) by the use of the ADAM optimization algorithm as implemented in tensorflow 1.13 (Abadi et al., 2016). To prevent overfitting, we applied dropout regularization to the different network layers (Srivastava et al., 2014) and global gradient norm clipping (with a clipping threshold of 5; Pascanu et al., 2013). During the training, we set the dropout probability to 0% for the first four convolution layers, 20% for the next four convolution layers, and 40% for the last four convolution layers. Each training epoch was defined as a complete iteration over all samples in the respective training dataset. We used a learning rate of 0.0001 and a batch size of 32. All network weights were initialized by the use of a normal-distributed random initialization scheme (Glorot and Bengio, 2010). The DL model was written in tensorflow 1.13 (Abadi et al., 2016) and the interprettensor toolbox (https://github.com/VigneshSrinivasan10/interprettensor).

3D-DeepLight We iteratively trained the 3D-DeepLight architecture through back-propagation (Hecht-Nielsen, 1992) by the use of the ADAM optimization algorithm as implemented in tensorflow 1.13 (Abadi et al., 2016). To prevent overfitting, we applied dropout regularization to the convolutional network layers (Srivastava et al., 2014), by setting the dropout probability to 20% during training. Each training epoch was defined as a complete iteration over all samples in the respective training dataset. We used a learning rate of 0.001 and a batch size of 32. All network weights were initialized by the use of a normal-distributed random initialization scheme (Glorot and Bengio, 2010). The DL model was written in tensorflow 1.13 (Abadi et al., 2016) and the iNNvestigate toolbox (Alber et al., 2019).

B.1.8 GLM analysis details for study II

FMRI data: Our GLM subject-level analyses of the fMRI data included one predictor for each of the four cognitive states in the design matrix (each representing a box-car function for the occurrence of a cognitive state; for methodological details on the GLM, see section 2.3.2). We convolved these predictors with a canonical glover haemodynamic response function (HRF; Lindquist et al., 2009) as implemented in Nistats 0.0.1rc0 (`nistats.github.io/`), to generate the model predictors. We added temporal derivative terms derived from each predictor, an intercept and an indicator of the experiment run to the design matrix, which we all treated as confounds of no interest. The

derivative terms were computed by the use of the cosine drift model as implemented in Nistats 0.0.1rc0 (`nistats.github.io/`).

To generate a set of group-level brain maps with the GLM, we computed a second-level GLM contrast by the use of the standard two-stage procedure for a random-effects group-level analysis, as proposed by Holmes and Friston (1998). Here, the subject-level regression coefficients β are treated as random effects in a second-level linear contrast analysis, where the distribution of first-level β-contrasts is assessed. Contrasts were computed between each cognitive state and all others. The resulting group-level brain maps show the Z-values resulting from this test.

Relevance data: Our GLM analyses of the relevance data resulting from the application of the LRP technique to DeepLight's decoding decisions (for an overview of the LRP technique, see section 2.3.3 of the main text) included one predictor for each of the four cognitive states in the data (each representing a box-car function for the occurrence of a cognitive state). Our previous analyses have indicated that the relevance data show a similar temporal evolution as the HRF (see Fig. 2.4 of the main text). For this reason, we next convolved the predictors with a canonical glover HRF (Lindquist et al., 2009), as implemented in Nistats 0.0.1rc0 (`nistats.github.io/`), to generate a set of model predictors.

We further added temporal derivative terms derived from each predictor, an intercept and an indicator of the experiment run to the design matrix. The temporal derivative terms were computed by the use of the cosine drift model as implemented in Nistats 0.0.1rc0 (`nistats.github.io/`). Additionally, we added one regressor to the design matrix, indicating the total sum of relevance values for each fMRI volume, to account for the variability in the sum of relevance values between fMRI volumes (resulting from the different output predictions of the DL model; for an overview of the LRP technique, see section 2.3.3 of the main text). To also account for non-linear relationships between this regressor and the relevance values, we added a regressor for the first derivative of the relevance sums, the squared relevance sums, and the first derivative of the squared relevance sums to the design matrix. All of these predictors were treated as confounds of no interest.

Lastly, we added two regressors to the design matrix indicating whether DeepLight correctly or incorrectly identified the cognitive state of each fMRI volume (again in form of two inverse box-car functions). Importantly, we accounted for these two predictors in each computed contrast, by contrasting each cognitive states against all others and corrected versus incorrectd decisions (e.g., if the wanted to compute a contrast for the body state, we would set the contrast vector to: 3, -1, -1, -1, 1, -1, for the predictors: body, face, place, tool, correct, incorrect).

To generate a set of group-level brain maps with the GLM, we computed a second-level GLM contrast by the use of the standard two-stage procedure for a random-effects group-level analysis, as proposed by Holmes and Friston (1998). Here, the subject-level regression coefficients β are treated as random effects in a second-level linear contrast analysis, where the distribution of first-level β-contrasts is assessed. The resulting group-level brain maps show the Z-values resulting from this test.

B.1.9 NeuroSynth

The goal of NeuroSynth (Yarkoni et al., 2011) is to provide an automated meta-analysis database relating cognitive states and brain activity. For specific cognitive states (i.e., "pain"), the NeuroSynth database incorporates a large record of neuroimaging studies that used this term at a high frequency (> 1 in 1000 words) in the article text. For these studies, the database includes the activation coordinates from all tables that are reported in these studies, producing a large set of *term-to-activation* mappings. Based on these mappings, NeuroSynth provides two types of tests: a *uniformity test*, indicating whether the probability that an article reports a specific brain activation is different, if it includes a specific term, compared to when brain activation would be distributed uniformly throughout gray matter and an *association test*, indicating whether the probability that a research article reports a specific brain activation is different, if it includes a specific term, compared to when it does not.

For our analyses, we used the latter, association test, as recommended by the NeuroSynth authors (Yarkoni et al., 2011), and extracted the thresholded ($P \leqslant 0.01$, voxel-wise false discovery rate corrected Benjamini and Hochberg, 1995) brain maps for the four stimulus classes (indicated by the terms "body", "face", "place" and "tools"). These maps indicate a Z-value for the previously described association test at each coordinate in the MNI-space.

B.1.10 F1-score

The F1-score for a binary classifier and a given dataset is defined as the harmonic mean of its precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{B.1}$$

Here, the classifier's precision is defined as the fraction of samples in the dataset that it correctly classified as positives, given the total number of samples that it classified as positive in the dataset, whereas its recall describes the fraction of samples that it correctly classified as positive, given the overall number of positive samples in the dataset.

B.1.11 fMRIPrep details for Multi-task data

This dataset was processed using *fMRIPrep* 20.0.5 (Esteban et al. (2019); Esteban et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502). For

Anatomical data preprocessing The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue

segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

Functional data preprocessing For each of the 18 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using `flirt` (FSL 5.0.9, Jenkinson and Smith, 2001) with the boundary-based registration (Greve and Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). BOLD runs were slice-time corrected using `3dTshift` from AFNI 20160207 (Cox and Hyde, 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007a). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM

regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

Copyright Waiver The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license.

B.1.12 fMRIPrep details for HCP working memory task

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.0.5 (Esteban et al. (2019); Esteban et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.4.2 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9,

RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to two standard spaces (MNI152NLin6Asym, MNI152NLin2009cAsym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym], *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym],

Functional data preprocessing For each of the 14 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using `flirt` (FSL 5.0.9, Jenkinson and Smith, 2001) with the boundary-based registration (Greve and Fischl, 2009) cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin6Asym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007a). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation).

Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

B.2 Figures

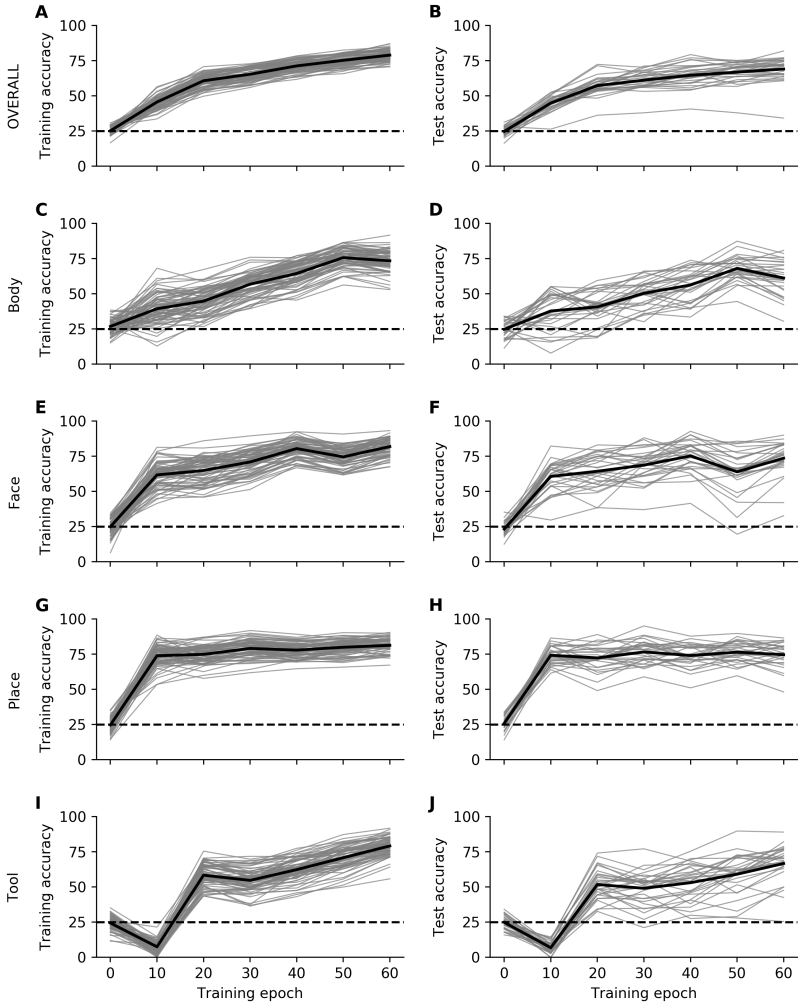

Figure B.1: DeepLight's decoding accuracy as a function of the training epochs in the data of the HCP working memory task of study I. A-B: Overall decoding accuracy achieved in the training (A) and test (B) data. Thick black lines indicate the grand average, whereas thin grey lines indicate individual subjects. Decoding accuracy in the training and test data for the body (C-D), face (E-F), place (G-H) and tool (I-J) stimulus classes. An epoch is defined as a full iteration over the training data. We define decoding accuracy as the fraction of samples in the data that were classified correctly.

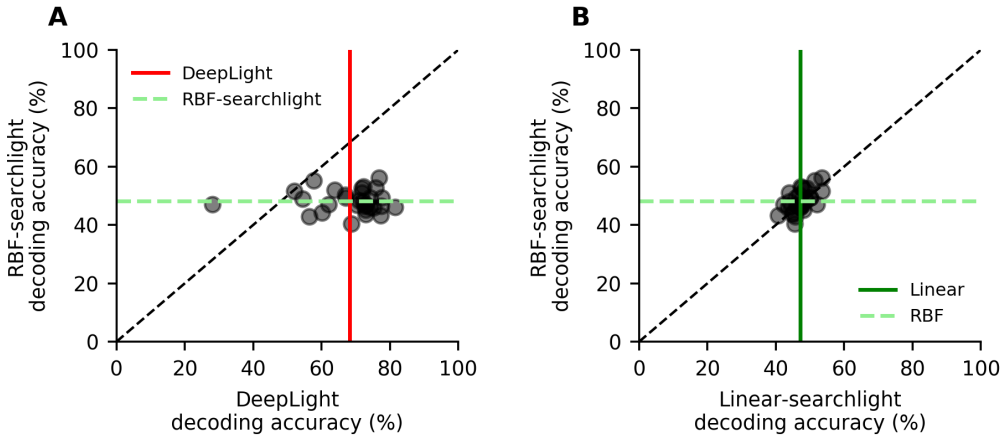

Figure B.2: Out-of-sample decoding performance of the searchlight analysis with a linear-kernel SVM and a non-linear radial basis function (RBF) kernel SVM in the test data of the HCP working memory task used in study I. We trained each of the two searchlight variants on the data of the first experiment run of a subject, before predicting the cognitive state for each TR of the second experiment run (for details on the estimation procedures, see Appendix B.1.5). We performed this prediction exercise only within the data of the subjects in the test dataset. We fixed the searchlight radius to 5.6mm, while we set γ parameter of the RBF-kernel to 1 across all subjects. A: Decoding performance comparison of the RBF-kernel SVM with 2D-DeepLight. B: Decoding performance comparison of the RBF-kernel SVM with the linear-kernel SVM. Black points indicate average decoding accuracies of individual subjects. Colored lines indicate averages across subjects. For an overview of the statistical results of the comparison, see section 2.4.1 of the main text.

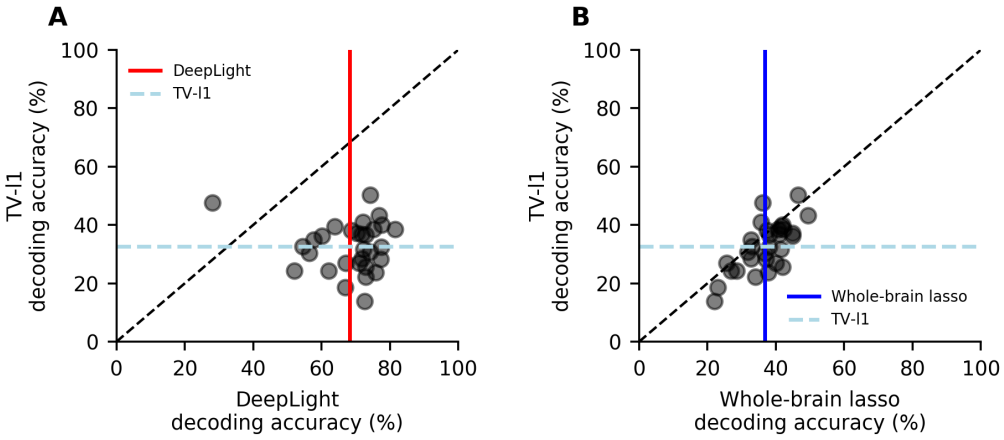

Figure B.3: Out-of-sample decoding performance of the whole-brain lasso and its TV-L1 extension in the test data of the HCP working memory task used in study I. TV-L1 extends the logistic regression model of the whole-brain lasso by combining the L1-penalty of the whole-brain lasso with an additional Total-Variation (TV) penalty to better account for the spatial dependency structure of fMRI data (for details on the TV-L1 approach, see (Gramfort et al., 2013)). The trade-off between both penalty terms is determined by a mixing constant (the L1-ratio, bounded between $[0, 1]$), which determines the ratio at which both penalty terms are mixed during regularization (with larger L1-ratios indicating stronger L1-regularization). To estimate the parameters of both decoding approaches, we first trained both on the data of the first experiment run of a subject (for details on the subject-level estimation procedures of the whole-brain lasso, see Appendix B.1.5). Subsequently, we used the trained decoding models to predict the cognitive state underlying each fMRI volume of the second experiment run of the same subject. We performed this prediction exercise only within the data of the 30 subjects in the held-out test dataset. To determine the best fitting L1-ratio of TV-L1 for each subject, we evaluated five different L1-ratios (namely, 0.1, 0.3, 0.5, 0.7, 0.9) in a 3-fold cross-validation procedure within the data of the first experiment run of a subject. We then selected the L1-ratio achieving the highest decoding accuracy across the three folds. The overall strength of the regularization (as determined by the λ-parameter, see section 2.3.2 of the main text), was determined through the default grid-search procedure implemented by the Nilearn toolbox (Abraham et al., 2014). Here, an additional grid of 10 distinct λ-values is evaluated in the same 3-fold cross-validation procedure described before. A: Decoding performance comparison of TV-L1 with 2D-DeepLight. B: Decoding performance comparison of TV-L1 with the whole-brain lasso. Each point represents an individual subject. Colored lines indicate averages across subjects. For an overview of the statistical results of the comparison, see section 2.4.1 of the main text.

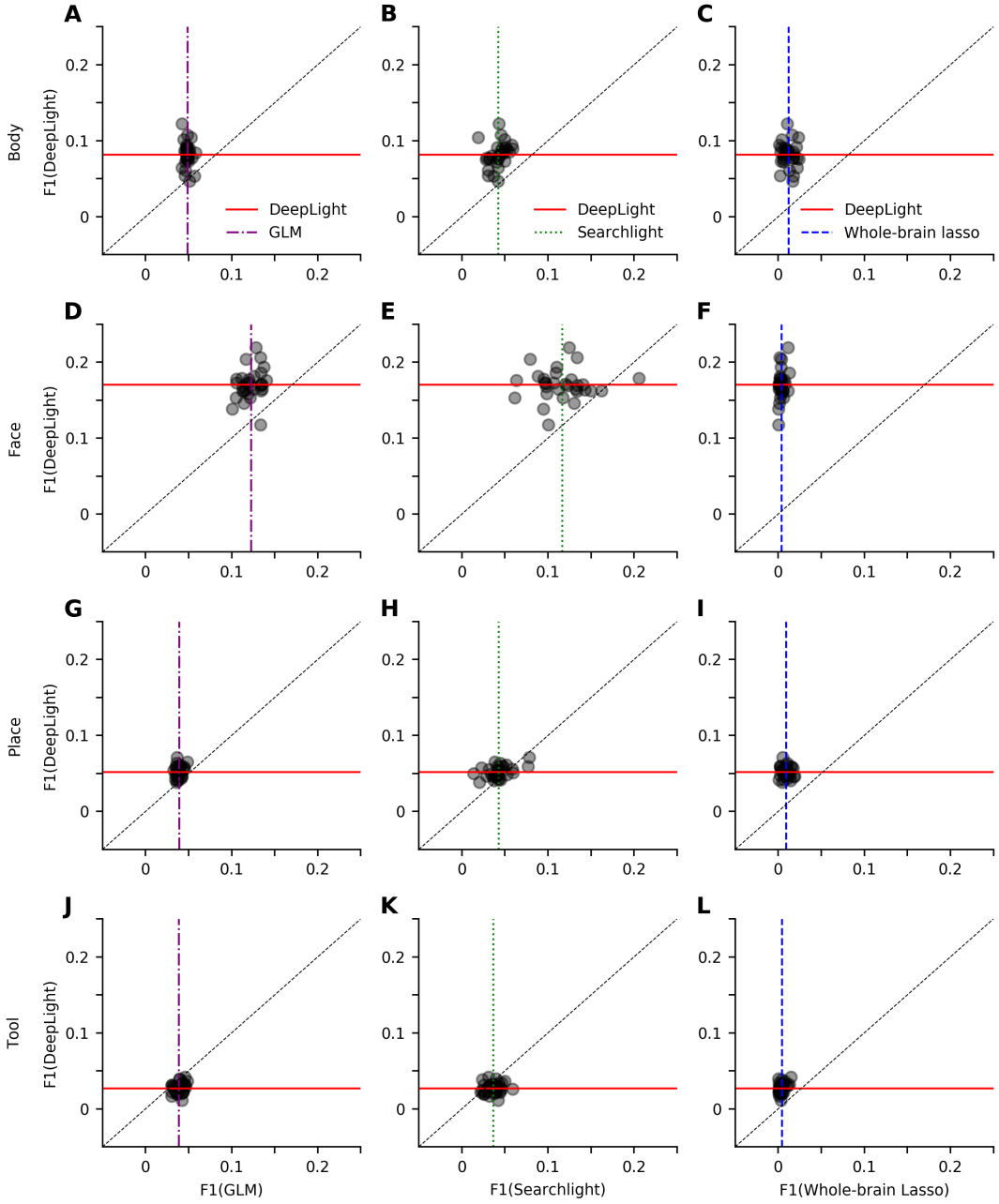

Figure B.4: Comparison of the subject-level F1-scores of 2D-DeepLight to the GLM (A,D,G,J), searchlight analysis (B,E,H,K) and whole-brain lasso (C,F,I,L) in the test data of the HCP working memory task used in study I. The brain maps of DeepLight, the searchlight analysis, and whole-brain lasso were thresholded at the 90th percentile of their values, whereas the GLM brain maps were thresholded at a P-value of 0.005 (uncorrected). Scatter points indicate the F1-scores of individual subjects. Colored lines indicate average F1-scores across subjects. For an overview of the statistical results of the comparison, see section 2.4 of the main text.

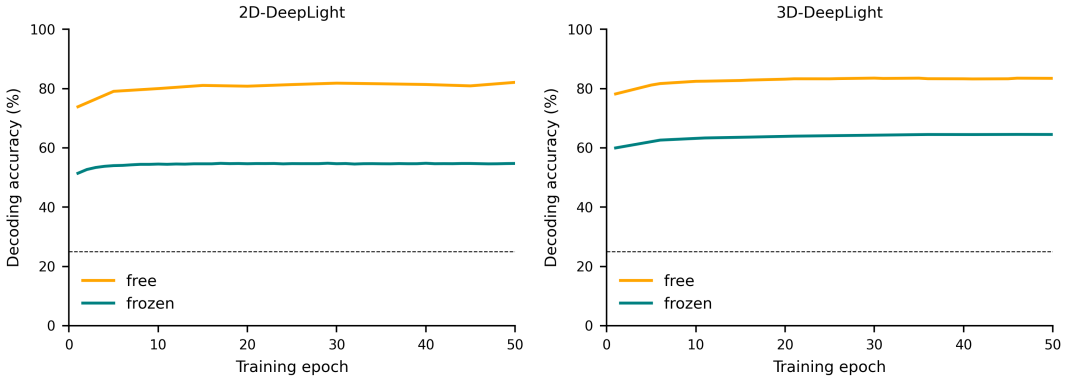

Figure B.5: Comparison of two different fine-tuning approaches in the validation data of the HCP working memory task of study II (see section 2.3.1 of the main text). We initialized the weights of two variants of each architecture (left: 2D-DeepLight, right: 3D-DeepLight) to the weights of the pre-trained models (all except for the output unit, which now included four instead of 16 neurons; see section 2.3.3 of the main text for an overview of the architectures and Fig. 2.7 of the main text for an overview of the pre-trained models). For one variant of each architecture we froze the pre-trained weights during fine-tuning (depicted in green), while the other model variant was allowed to train all of its weights during fine-tuning (depicted in yellow) (see Appendix B.1.7 for an overview of the training procedures). Lines indicate decoding accuracy in the validation data as a function of the training epochs.

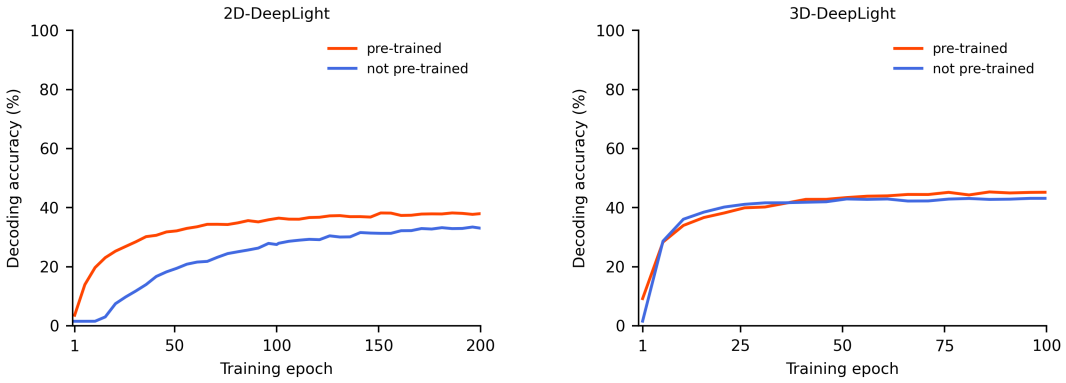

Figure B.6: Training decoding accuracy for a pre-trained and not pre-trained variant of the 2D- (left) and 3D- (right) DeepLight architecture in the validation data of the "Multi-task" dataset (see section 2.3.1 of the main text; see Appendix B.1.7 for an overview of the training procedures). An epoch was defined as an entire iteration over the training dataset. Lines indicate decoding accuracy. Colors indicate the pre-trained (red) and not pre-trained (blue) model variants.

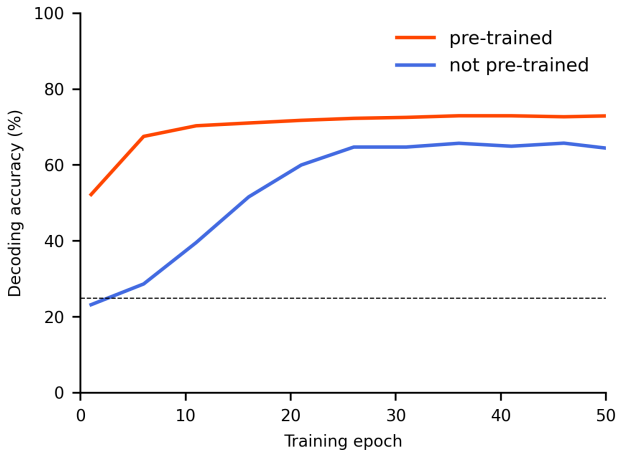

Figure B.7: Training decoding accuracy for a pre-trained (red) and not pre-trained (blue) 3D-DeepLight variant in the validation data of the HCP working memory task used in study II that was preprocessed with fMRIPrep (see section 2.4 of the main text and Appendix B.1.12; see Appendix B.1.7 for an overview of the training procedures). An epoch was defined as an entire iteration over the training dataset. Lines indicate decoding accuracy.

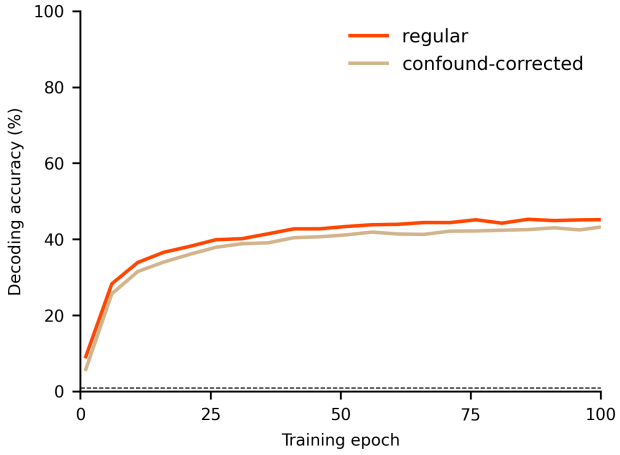

Figure B.8: Training decoding accuracy for the pre-trained 3D-DeepLight variant in two conditions: when it is fine-tuned on the regular fMRI data of the Multi-task dataset (red) or on a version that is corrected for basic noise confounds (tan). Specifically, we corrected the Mutli-task data for any variance resulting from the six parameters of basic motion correction, as well as the three temporal and anatomical noise components with the largest singular values resulting from fMRIPrep's CompCor method (for details on this methods, see (Behzadi et al., 2007b)), by regressing their variance out of the time-series signal of each voxels (as implemented in Nilearns "signal.clean" function Abraham et al., 2014). See Appendix B.1.7) for an overview of the training procedures. An epoch was defined as an entire iteration over the training dataset. Lines indicate decoding accuracy.

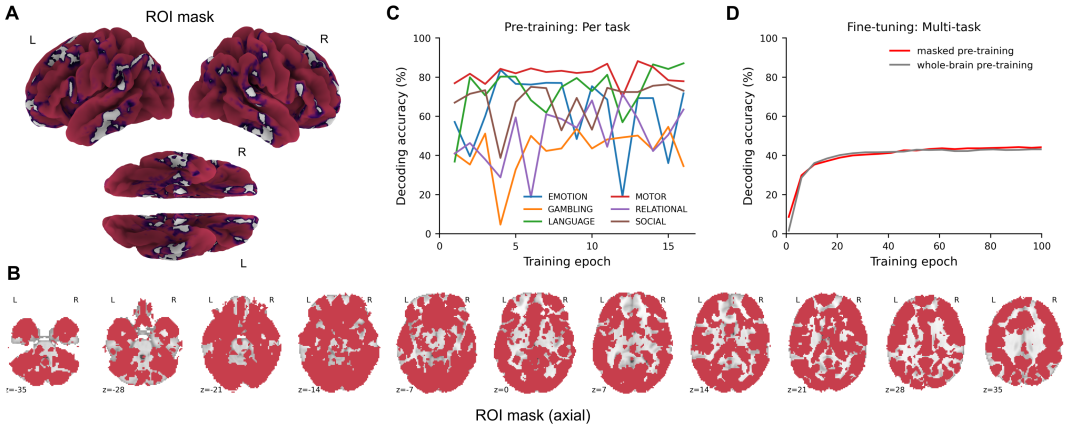

Figure B.9: Test of a 3D-DeepLight variant that was pre-trained on a masked version of the pre-training dataset. A-B: Brain mask for the fMRI data of the six HCP tasks that we use for pre-training (A, B; see section 2.3.1 of the main text for an overview of the data). We computed the mask by performing a two-stage GLM analysis for 50 randomly selected subjects in each task (for details on the GLM analysis, see Appendix B.1.8) and thresholding the resulting brainmaps at a false-discovery rate of 0.1. C: Training decoding accuracy of the 3D-DeepLight variant in the validation data of the masked pre-training dataset (including 50 subjects per task; see section 2.4.2 of the main text). Colors indicate the different HCP pre-training tasks. D: Training decoding accuracy in the HCP working memory task of the pre-trained 3D-DeepLight variants that were either trained with the masked (red) or whole-brain (grey) pre-training dataset. Lines indicate decoding accuracy as a function of training epochs.

B.3 Tables

Test Subject	C	Decoding Accuracy
1	18.74	33%
2	0.26	33%
3	0.09	42%
4	0.66	42%
5	0.22	47%
6	0.26	49%
7	0.79	37%
8	22.57	42%
9	12.92	26%
10	83.02	40%
11	39.44	27%
12	22.57	37%
13	39.44	32%
14	0.26	42%
15	0.79	38%
16	3.51	38%
17	3.51	41%
18	0.26	22%
19	0.15	36%
20	27.19	38%
21	32.75	45%
22	7.39	34%
23	6.14	41%
24	7.39	45%
25	0.66	29%
26	4.23	38%
27	2.92	33%
28	47.51	42%
29	12.92	36%
30	1.67	23%

Table B.1: Selected regularization strength parameters for the subject-level whole-brain lasso analysis (for details on the underlying subject-level grid search procedure, see Appendix B.1.5). For each subject, the selected regularization parameter (C) and resulting decoding accuracy are presented.

λ	Decoding Accuracy
1e-7	46.62%
1e-6	46.40%
1e-5	47.82%
0.0001	**48.11%**
0.0002	47.24%
0.0003	47.42%
0.0005	46.11%
0.0008	47.03%
0.001	45.53%
0.002	45.58%
0.004	44.98%
0.007	44.86%
0.01	42.89%
0.02	40.69%
0.03	37.14%
0.06	35.48%
0.1	26.57%
0.2	25.13%
0.3	25.00%
0.5	25.09%

Table B.2: λ parameters of the group-level whole-brain lasso anlaysis that were evaluated in the grid search procedure (for details on the group-level grid search, see Appendix B.1.5). The respective decoding accuracy in the test dataset is given for each λ value.

P	Percentile	GLM	Searchlight	Whole-brain lasso
0.05	85	t(29)=8.54, $p < 0.0001$	t(29)=11.99, $p < 0.0001$	t(29)=21.64, $p < 0.0001$
0.05	90	t(29)=10.46, $p < 0.0001$	t(29)=13.26, $p < 0.0001$	t(29)=20.93, $p < 0.0001$
0.05	95	t(29)=11.43, $p < 0.0001$	t(29)=14.04, $p < 0.0001$	t(29)=18.41, $p < 0.0001$
0.005	85	t(29)=8.54, $p < 0.0001$	t(29)=11.99, $p < 0.0001$	t(29)=21.61, $p < 0.0001$
0.005	90	t(29)=10.46, $p < 0.0001$	t(29)=13.26, $p < 0.0001$	t(29)=20.93, $p < 0.0001$
0.005	95	t(29)=11.43, $p < 0.0001$	t(29)=14.04, $p < 0.0001$	t(29)=18.41, $p < 0.0001$
0.0005	85	t(29)=8.54, $p < 0.0001$	t(29)=11.99, $p < 0.0001$	t(29)=21.61, $p < 0.0001$
0.0005	90	t(29)=10.45, $p < 0.0001$	t(29)=13.26, $p < 0.0001$	t(29)=20.93, $p < 0.0001$
0.0005	95	t(29)=11.42, $p < 0.0001$	t(29)=14.04, $p < 0.0001$	t(29)=18.41, $p < 0.0001$
0.00005	85	t(29)=8.54, $p < 0.0001$	t(29)=11.99, $p < 0.0001$	t(29)=21.61, $p < 0.0001$
0.00005	90	t(29)=10.46, $p < 0.0001$	t(29)=13.26, $p < 0.0001$	t(29)=20.93, $p < 0.0001$
0.00005	95	t(29)=11.42, $p < 0.0001$	t(29)=14.04, $p < 0.0001$	t(29)=18.41, $p < 0.0001$

Table B.3: Results of a two-sided t-test of the difference between the subject-level F1-scores for the body stimulus class of each of the competing analysis approaches with the subject-level F1-scores of DeepLight. The t-test comparison was repeated for each combination of percentile- and P-threshold (for details on the F1-score comparison procedure, see section 2.4.1 of the main text). T-tests were performed using a Bonferroni adjusted alpha level of 0.0014 (0.05/36). Bold font indicates t-tests with p-values greater than 0.0014.

P	Percentile	GLM	Searchlight	Whole-brain lasso
0.05	85	t(29)=16.26, $p < 0.0001$	t(29)=9.23, $p < 0.0001$	t(29)=61.12, $p < 0.0001$
0.05	90	t(29)=13.04, $p < 0.0001$	t(29)=8.57, $p < 0.0001$	t(29)=48.32, $p < 0.0001$
0.05	95	t(29)=4.97, p=0.00027	t(29)=7.01, $p < 0.0001$	t(29)=33.92, $p < 0.0001$
0.005	85	t(29)=16.26, $p < 0.0001$	t(29)=9.23, $p < 0.0001$	t(29)=61.12, $p < 0.0001$
0.005	90	t(29)=13.04, $p < 0.0001$	t(29)=8.57, $p < 0.0001$	t(29)=48.32, $p < 0.0001$
0.005	95	t(29)=4.97, p=0.00027	t(29)=7.01, $p < 0.0001$	t(29)=33.92, $p < 0.0001$
0.0005	85	t(29)=16.26, $p < 0.0001$	t(29)=9.23, $p < 0.0001$	t(29)=61.12, $p < 0.0001$
0.0005	90	t(29)=13.04, $p < 0.0001$	t(29)=8.57, $p < 0.0001$	t(29)=48.32, $p < 0.0001$
0.0005	95	t(29)=4.97, p=0.00027	t(29)=7.01, $p < 0.0001$	t(29)=33.92, $p < 0.0001$
0.00005	85	t(29)=16.26, $p < 0.0001$	t(29)=9.23, $p < 0.0001$	t(29)=61.12, $p < 0.0001$
0.00005	90	t(29)=13.04, $p < 0.0001$	t(29)=8.57, $p < 0.0001$	t(29)=48.32, $p < 0.0001$
0.00005	95	t(29)=4.97, p=0.00027	t(29)=7.01, $p < 0.0001$	t(29)=33.92, $p < 0.0001$

Table B.4: Results of a two-sided t-test of the difference between the subject-level F1-scores for the face stimulus class of each of the competing analysis approaches with the subject-level F1-scores of DeepLight. The t-test comparison was repeated for each combination of percentile- and P-threshold (for details on the F1-score comparison procedure, se section 2.4.1 of the main text). T-tests were performed using a Bonferroni adjusted alpha level of 0.0014 (0.05/36). Bold font indicates t-tests with p-values greater than 0.0014.

P	Percentile	GLM	Searchlight	Whole-brain lasso
0.05	85	t(29)=5.02, p=0.00024	**t(29)=1.49, p=0.15**	t(29)=20.56, $p < 0.0001$
0.05	90	t(29)=9.26, $p < 0.0001$	t(29)=4.25, p=0.0002	t(29)=22.43, $p < 0.0001$
0.05	95	t(29)=11.87, $p < 0.0001$	t(29)=7.63, $p < 0.0001$	t(29)=22.38, $p < 0.0001$
0.005	85	t(29)=5.02, p=0.00024	**t(29)=1.49, p=0.15**	t(29)=20.56, $p < 0.0001$
0.005	90	t(29)=9.26, $p < 0.0001$	t(29)=4.25, p=0.0002	t(29)=22.43, $p < 0.0001$
0.005	95	t(29)=11.87, $p < 0.0001$	t(29)=7.63, $p < 0.0001$	t(29)=22.38, $p < 0.0001$
0.0005	85	t(29)=5.02, p=0.00024	**t(29)=1.49, p=0.15**	t(29)=20.56, $p < 0.0001$
0.0005	90	t(29)=9.26, $p < 0.0001$	t(29)=4.25, p=0.0002	t(29)=22.43, $p < 0.0001$
0.0005	95	t(29)=11.87, $p < 0.0001$	t(29)=7.63, $p < 0.0001$	t(29)=22.38, $p < 0.0001$
0.00005	85	t(29)=5.02, p=0.00024	**t(29)=1.49, p=0.15**	t(29)=20.56, $p < 0.0001$
0.00005	90	t(29)=9.26, $p < 0.0001$	t(29)=4.25, p=0.0002	t(29)=22.43, $p < 0.0001$
0.00005	95	t(29)=11.87, $p < 0.0001$	t(29)=7.63, $p < 0.0001$	t(29)=22.38, $p < 0.0001$

Table B.5: Results of a two-sided t-test of the difference between the subject-level F1-scores for the place stimulus class of each of the competing analysis approaches with the subject-level F1-scores of DeepLight. The t-test comparison was repeated for each combination of percentile- and P-threshold (for details on the F1-score comparison procedure, see the section 2.4.1 of the main text). T-tests were performed using a Bonferroni adjusted alpha level of 0.0014 (0.05/36). Bold font indicates t-tests with p-values greater than 0.0014.

P	Percentile	GLM	Searchlight	Whole-brain lasso
0.05	85	t(29)=-8.41, $p < 0.0001$	t(29)=-5.17, p=0.0002	t(29)=20.66, $p < 0.0001$
0.05	90	t(29)=-8.19, $p < 0.0001$	t(29)=-4.39, p=0.0001	t(29)=18.31, $p < 0.0001$
0.05	95	t(29)=-8.04, $p < 0.0001$	**t(29)=-3.10, p=0.004**	t(29)=15.06, $p < 0.0001$
0.005	85	t(29)=-8.42, $p < 0.0001$	t(29)=-5.17, p=0.0002	t(29)=20.66, $p < 0.0001$
0.005	90	t(29)=-8.19, $p < 0.0001$	t(29)=-4.39, p=0.0001	t(29)=18.31, $p < 0.0001$
0.005	95	t(29)=-8.04, $p < 0.0001$	**t(29)=-3.10, p=0.004**	t(29)=15.06, $p < 0.0001$
0.0005	85	t(29)=-8.42, $p < 0.0001$	t(29)=-5.17, p=0.0002	t(29)=20.66, $p < 0.0001$
0.0005	90	t(29)=-8.19, $p < 0.0001$	t(29)=-4.39, p=0.0001	t(29)=18.31, $p < 0.0001$
0.0005	95	t(29)=-8.04, $p < 0.0001$	**t(29)=-3.10, p=0.004**	t(29)=15.06, $p < 0.0001$
0.00005	85	t(29)=-8.42, $p < 0.0001$	t(29)=-5.17, p=0.0002	t(29)=20.66, $p < 0.0001$
0.00005	90	t(29)=-8.19, $p < 0.0001$	t(29)=-4.39, p=0.0001	t(29)=18.31, $p < 0.0001$
0.00005	95	t(29)=-8.04, $p < 0.0001$	**t(29)=-3.10, p=0.004**	t(29)=15.05, $p < 0.0001$

Table B.6: Results of a two-sided t-test of the difference between the subject-level F1-scores for the tool stimulus class of each of the competing analysis approaches with the subject-level F1-scores of DeepLight. The t-test comparison was repeated for each combination of percentile- and P-threshold (for details on the F1-score comparison procedure, see section 2.4.1 of the main text). T-tests were performed using a Bonferroni adjusted alpha level of 0.0014 (0.05/36). Bold font indicates t-tests with p-values greater than 0.0014.