Principles of Human Learning

Dissertation

zur Erlangung des Grades einer/eines

Doktorin/Doktor der Naturwissenschaften

(Dr. rer.nat.)

des Fachbereichs Psychologie der Philipps-Universität Marburg

Vorgelegt von

Marcel Binz, M.Sc.

Aus Oberwesel

Marburg, 11.01.2021

Vom Fachbereich Psychologie der Philipps-Universität Marburg
(Hochschulkennziffer 1180) als Dissertation angenommen am

Erstgutachter(in):	Prof. Dr. Dominik Endres
Zweitgutachter(in):	Dr. Anna Thorwart

Tag der Disputation: 20.04.2021

Ich versichere, dass ich die vorgelegte Dissertation selbst und ohne fremde Hilfe verfasst, nicht andere als die in ihr angegebenen Quellen oder Hilfsmittel benutzt, alle vollständig oder sinngemäß übernommenen Zitate als solche gekennzeichnet sowie die Dissertation in der vorliegenden oder einer ähnlichen Form noch bei keiner anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht habe.

(Unterschrift)

Principles of Human Learning

Abstract

What are the general principles that drive human learning in different situations? I argue that much of human learning can be understood with just three principles. These are *generalization*, *adaptation*, and *simplicity*. To verify this conjecture, I introduce a modeling framework based on the same principles. This framework combines the idea of *meta-learning* – also known as *learning-to-learn* – with the *minimum description length* principle. The models that result from this framework capture many aspects of human learning across different domains, including decision-making, associative learning, function learning, multi-task learning, and reinforcement learning. In the context of decision-making, they explain why different heuristic decision-making strategies emerge and how appropriate strategies are selected. The same models furthermore capture order effects found in associative learning, function learning and multi-task learning. In the reinforcement learning context, they resemble individual differences between human exploration strategies and explain empirical data better than any other strategy under consideration. The proposed modeling framework – together with its accompanying empirical evidence – may therefore be viewed as a first step towards the identification of a minimal set of principles from which all human behavior derives.

Prinzipien des Menschlichen Lernens

Zusammenfassung

Was sind die allgemeinen Prinzipien, die das menschliche Lernen in verschiedenen Situationen antreiben? Ich behaupte, dass ein Großteil des menschlichen Lernens mit nur drei Prinzipien verstanden werden kann. Diese sind *generalization*, *adaptation* und *simplicity*. Um diese Hypothese zu überprüfen, führe ich ein Modellierungsframework ein, das auf denselben Prinzipien basiert. Dieses Framework kombiniert die Idee des *meta-learning* – auch als *learning-to-learn* bekannt – mit dem *minimum description length* Prinzip. Die Modelle, die sich aus diesem Framework ergeben, erfassen viele Aspekte des menschlichen Lernens in verschiedenen Bereichen, einschließlich dem Entscheidungsfinden, dem assoziativem Lernen, dem Funktionslernen, dem Lernen mit mehreren Aufgaben und dem bestärkenden Lernen. Im Kontext der Entscheidungsfindung erklären sie, warum unterschiedliche heuristische Entscheidungsstrategien entstehen und wie geeignete Strategien ausgewählt werden. Dieselben Modelle erfassen außerdem Anordnungseffekte, die beim assoziativen Lernen, beim Funktionslernen und beim Lernen mit mehreren Aufgaben auftreten. Im Kontext des bestärkenden Lernens spiegeln sie individuelle Unterschiede zwischen menschlichen Explorationsstrategien wider und erklären empirische Daten besser als jede andere in Betracht gezogene Strategie. Das vorgeschlagene Modellierungsframework kann daher – zusammen mit den dazugehörigen empirischen Befunden – als erster Schritt zur Identifizierung einer minimalen Menge von Prinzipien, von denen das gesamte menschliche Verhalten abgeleitet werden kann, angesehen werden.

Acknowledgments

The fact that I ended up doing my PhD in Marburg was largely a big coincidence. I had already lined up a position in another lab, but the conditions there were not what I was hoping for, and so I decided to switch at the last minute. However, I could not have imagined a better place for spending my last three years than Marburg. This was at large due to the people that I got the chance to interact with.

First of all, there is my thesis adviser Dominik Endres, who was not only a wealth of knowledge about virtually any scientific topic but also gave me much freedom to pursue topics of my own interest. I was lucky to be part of such an amazing research group, whose members include Benjamin Knopp, Anja Kruse, Neda Meibodi, Hauke Niehaus, Mortiz Schubert, Raphael Schween, Christine Ulrich, Dmytro Velychko, as well as the countless HiWis and Bachelor/Master students. Each of you made everyday activities around the office more enjoyable in your own right. My thanks also go to all of the people involved in the GRK 2271 for creating a vibrant research environment. Especially to Anna Thorwart, my secondary advisor, who broadened my horizon in several discussions.

I am also grateful to Sam Gershman and his whole research group, who hosted me for three months at Harvard. Sam greatly shaped the way of how I think about psychology and about research in general. Eric Schulz, whom I met during this research visit, became a great inspiration for what I want to be as a scientist and as a person. This dissertation would only be half as good if not for the extensive and helpful feedback that Verena Heusser and Charley Wu provided. The final – and most important – thanks go to my parents for encouraging me to think critically, and to my brothers and friends for all the fun times.

Contents

Nomenclature

Numbers, Arrays and Sets

a	Scalar
$\mathbf{a}$	Vector
$\mathbf{A}$	Matrix
$\mathbf{I}$	Identity matrix
$\mathbb{R}$	The set of real numbers
$\{a, b\}$	The set of objects a and b
$a_{1:T}$	$a_1, a_2, \ldots, a_T$

Linear Algebra and Calculus

$\mathbf{a}^T$	Transpose of vector $\mathbf{a}$
$\mathbf{a} \odot \mathbf{b}$	Element-wise product of $\mathbf{a}$ and $\mathbf{b}$
$\nabla_{\mathbf{b}}\, a$	Gradient of a with respect to $\mathbf{b}$
$\int f(a)da$	Definite integral over the entire domain of a

Probability and Information Theory

$p(a)$	Probability distribution over a
$\mathbb{E}_{p(a)}\left[a\right]$	Expectation of a with respect to $p(a)$
$H\left[p(a)\right]$	Shannon entropy of the random variable a
$\mathrm{KL}\left[p(a)\vert\vert q(a)\right]$	Kullback-Leibler divergence between p and q
$\mathcal{N}(\mathbf{a}\vert\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution over $\mathbf{a}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

$\mathcal{U}(\mathbf{a}\|b, c)$	Uniform distribution over $\mathbf{a}$ with minimum b and maximum c

Functions

$\log$	Natural logarithm
$\log_2$	Logarithm to base 2
σ	Logistic sigmoid function
$\tanh$	Hyperbolic tangent function
$\mathbf{\Phi}$	Cumulative distribution function of the standard normal distribution

Parameters

α	Learning rate
β	Regularization factor in BMI and RL^3
$\mathbf{\Theta}$	Parameters of a neural network
$\mathbf{\Lambda}$	Parameters of the encoding distribution
$\mathbf{w}$	Parameters of a linear model

Abbreviations

BMI	Bounded meta-learned inference
EW	Equal weighting
GRU	Gated recurrent unit
IO	Ideal observer
KL	Kullback–Leibler divergence
MCPL	Multiple-cue probability learning
MDP	Markov decision process
MI	Meta-learned inference
PXP	Protected exeedance probability
RNN	Recurrent neural network

SC	Single cue
TS	Thompson sampling
TTB	Take-the-best
UCB	Upper confidence bounds
VD	Value-directed exploration
WADD	Weighted addtive strategy

1

Introduction

1.1 Principles of Human Learning

From the hobby gardener who tries to figure out how to grow vegetables in his garden to someone attempting to master the game of chess, we constantly find ourselves in situations where we have to learn something new. You could even argue that without the ability to learn, we would not be able to get much done at all. But, how *do* people learn? How does a hobby gardener figure out what it takes to grow vegetables, and how does one become an expert at playing chess?

The main goal of this thesis is to identify general principles that drive human learning in different contexts. More specifically, I seek to establish a minimal set of principles from which all of human behavior derives. Clearly, this is a gigantic task, and I do not claim to solve it in its entirety. However, I hope that the results presented here can serve as the first step towards this goal. In particular, I put forward the following three principles (also see Figure 1.1):

- **Generalization**: People learn to make good inferences in new situations.
- **Adaptation**: How people learn is adapted to their environment.
- **Simplicity**: People use short and simple learning algorithms.

The roots of these principles date back to Herbert Simon's ideas on bounded rationality (Simon, 1956, 1990a). Simon emphasized – using his now-famous analogy of the two blades of a scissor – that models of human cognition need to take both the structure of the environment (i.e., adaptation) and cognitive limitations of the mind (i.e., simplicity) into account (Simon, 1990b). Ever since then, all three principles have received a significant amount of attention within psychology.

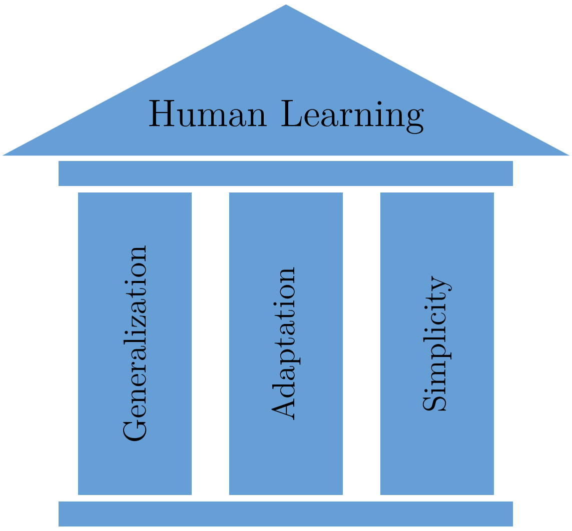


Figure 1.1: The three suggested principles of human learning: generalization, adaptation and simplicity.

From an empirical perspective, people show remarkable generalization abilities (Shepard, 1957, Shanks and Darby, 1998, Ghirlanda and Enquist, 2003). They can extrapolate well beyond the training distribution (DeLosh et al., 1997, Schulz et al., 2016b) and learn from just a few examples (Carey and Bartlett, 1978, Markman, 1989, Lake et al., 2015, 2017). Shepard (1987) – a pioneer in the studies of human generalization – even went as far as suggesting that "psychology's first general law should be a law of generalization" and subsequently proposed a theory of generalization that explains how people generalize from one stimulus to another based on an exponentially decaying function of the distance between the two stimuli.

The no free lunch theorem states that when averaged over all possible problems, no learning algorithm is better than another (Wolpert, 1996, Wolpert and

Macready, 1997). Therefore, learning algorithms need to be adapted to environment they are applied in. In his seminal work on rational analysis Anderson (1991b) asked “is human cognition adaptive?” and found evidence for this idea in memory retrieval, categorization, causal inference, and problem-solving. From an evolutionary perspective, this should come as no surprise, as we expect nature to select for individuals that perform well within their environment (Todd and Gigerenzer, 2007, 2012).

A general theory of human learning should also reflect that the brain only has a limited capacity (Simon, 1990a, Gershman et al., 2015, Lieder and Griffiths, 2020). Therefore, people need to use simple algorithms. The complexity of an algorithm can be defined as the shortest computer program that implements it (Kolmogorov, 1965, Solomonoff, 1964, Chaitin, 1969). In this view, simplicity is essentially compressibility (Feldman, 2016). Chaitin (2002) eloquently highlighted the importance of compression within intelligent systems by proclaiming that “comprehension is compression”. Echoes of this idea are found throughout all of cognitive science and psychology (Chater and Vitányi, 2003, Maguire et al., 2015, Feldman, 2000).

1.2 From Principles To Computational Models

Viewed independently all three principles are quite uncontroversial. But, how can we demonstrate that people actually follow them? The approach I take in this thesis is to build computational models that follow the same principles and then verify that these models do things that are similar to what people are doing. My main contribution is to provide implementations of learning algorithms that embody the three suggested principles and the following empirical investigation of how far we can get towards a general theory of human learning with a combination of them. The proposed framework relies on two key ideas: *meta-learning* (Bengio et al., 1990, Schmidhuber et al., 1996, Thrun and Pratt, 1998) and the *minimum description length* principle (Rissanen, 1978, Grünwald and Grunwald, 2007, Hinton and Van Camp, 1993).

Meta-learning refers to the concept of learning a learning algorithm based on previous experience. The meta-learning models used in this thesis accomplish this by parameterizing learning algorithms through *deep neural networks* (Goodfellow et al., 2016), which are then trained to make optimal inferences within a specific environment. Eventually, this leads to the emergence of an adapted learning algorithm that generalizes optimally to future data-points. Therefore, meta-learning covers two of the suggested principles; generalization and adaptation.

The last principle is realized by limiting the description length of the emerging learning algorithm through an additional information-theoretic regularization term. The description length of an algorithm is defined as the number of bits required to implement it. Limiting it, therefore, acts as a particular notion of simplicity.

Putting these two ideas together leads to a class of algorithms for solving supervised learning problems called *bounded meta-learned inference* (BMI) and to one for reinforcement learning problems called RL^3. By applying BMI and RL^3 to a set of diverse domains, I demonstrate that generalization, adaptation, and simplicity capture many characteristics of human learning across different research areas.

1.3 Summary

Chapter 2 starts with a review of computational theories of human learning. Therein, I also discuss how prior work has addressed generalization, adaptation, and simplicity. Chapter 3 introduces BMI and RL^3 in detail and contrasts them with other theories of learning. Chapters 4 to 6 then examine BMI and RL^3 on a diverse set of learning problems:

- Chapter 4 demonstrates that different heuristics, that have been previously suggested as models of human decision-making, emerge naturally from BMI. BMI furthermore makes precise predictions about if and when these heuristics should emerge, which were verified in three new experimental studies.
- Chapter 5 demonstrates that BMI additionally captures different order effects found in associative learning, function learning, and multi-task learning.
- Chapter 6 demonstrates that RL^3 discovers a diverse spectrum of exploration strategies that align with individual differences in human exploration on a two-armed bandit task.

In Chapter 7, I wrap up this thesis by discussing limitations of the proposed framework and suggest directions for future research.

2

Previous Theories of Human Learning

Do current computational theories of human learning already address the three suggested principles? This chapter reviews prevalent theories, with a focus on how they address generalization, adaptation, and simplicity. Table 2.1 provides a summary of all reviewed theories along with their characteristics.

The main focus of this chapter is on supervised learning problems. In such problems, an agent – either human or machine – has to learn how to map an input variable $\mathbf{x} \in \mathbb{R}^d$ to a target variable y. If the target variable is an unconstrained real number, this is known as a regression problem. If it belongs to a discrete set of categories, then this is a classification problem. I assume that data arrives sequentially and denote a sequence of input-target pairs of length t as $\mathbf{x}_{1:t}, y_{1:t}$. In each time-step, the agent observes an input. It then makes a prediction for that input and subsequently receives feedback about the actual target variable. This problem formulation is extremely general, and it maps directly to many experimental paradigms from the psychology literature, including those found in associative learning (Shanks, 1995), category learning (Ashby and Maddox, 2005) and decision-making (Gigerenzer and Gaissmaier, 2011). For example, in an associative learning setting $\mathbf{x}$ may correspond to the presence or absence of a stimulus and y to an associated reward, or in a decision-making setting $\mathbf{x}$ may represent different features of two football teams, while y indicates the outcome of a match between the two.

Describing how computational theories address generalization, adaptation, and simplicity requires formal definitions of the three principles. Therefore, I will discuss briefly how each of them can be formalized before relating them to different theories.

Framework	**Generalization**	**Adaptation**	**Simplicity**
Rescorla-Wagner	🟠	🔴	🔴
Bayesian Inference	🟢	🟢	🔴
Rational Process Models	🟢	🟢	🟠
Program Induction	🟢	🟢	🔴
Connectionism	🟠	🔴	🔴
Gradient-Based Meta-Learning	🟠	🟢	🔴
Model-Based Meta-Learning	🟢	🟢	🔴

Table 2.1: Properties of different learning frameworks. Green dots indicate that the corresponding principle is addressed optimally within the framework. Red dots indicate no consideration at all. Instances in between both ends are labeled with an orange dot.

Generalization

Being able to generalize means to perform well in novel situations. In machine learning, performance is typically measured in terms of a loss function. Let $\mathcal{L}(\mathbf{x}, y, \mathbf{x}_{1:t}, y_{1:t}, m)$ be such a loss function that compares predictions of model m on input $\mathbf{x}$ with the the target variable y after having observed t input-target examples $\mathbf{x}_{1:t}, y_{1:t}$. The generalization loss $\mathcal{L}_g$ is obtained by averaging $\mathcal{L}$ over all possible future data-points:

$$\mathcal{L}_g(\mathbf{x}_{1:t}, y_{1:t}, m) = \mathbb{E}_{p(\mathbf{x},y)}\left[\mathcal{L}(\mathbf{x}, y, \mathbf{x}_{1:t}, y_{1:t}, m)\right] \tag{2.1}$$

I say that a learning algorithm generalizes optimally if its generalization loss is as low as possible.

Adaptation

Adaptation can be defined in a similar fashion. It is the ability to adjust to an environment in a way that one learns well on tasks that are typically encountered within that particular environment. I define an environment as the distribution over tasks $p(\mathbf{x}_{1:t}, y_{1:t})$ can be encountered, and say that a learning algorithm is adapted to its environment if it minimizes performance averaged over all tasks:

$$\begin{aligned} \mathcal{L}_a(m) &= \mathbb{E}_{p(\mathbf{x}_{1:t},y_{1:t})}\left[\mathcal{L}_g(\mathbf{x}_{1:t}, y_{1:t}, m)\right] && (2.2) \\ &= \mathbb{E}_{p(\mathbf{x}_{1:t},y_{1:t})}\left[\mathbb{E}_{p(\mathbf{x},y)}\left[\mathcal{L}(\mathbf{x}, y, \mathbf{x}_{1:t}, y_{1:t}, m)\right]\right] && (2.3) \end{aligned}$$

Simplicity

Formally, an algorithm's complexity can be defined as the length of the shortest computer program that implements it (Kolmogorov, 1965, Solomonoff, 1964, Chaitin, 1969). Learning algorithms with low complexity are simple. Typically, simplicity is at odds with measures of performance such as generalization; one has to give away performance for simplicity. I say that a learning algorithm optimally trades-off performance for simplicity if no shorter algorithm that achieves the same level of performance exists.

2.1 Rescorla-Wagner Model

Historically, the *Rescorla-Wagner model* (Rescorla and Wagner, 1972) provides one of the earliest computational theories of human learning. The model assumes that predictions are computed through a linear combination between inputs and their corresponding associative strength $\mathbf{w}_t \in \mathbb{R}^d$. Learning – i.e., updating of associative strengths – is realized through gradient descent on a squared error loss function between the prediction and the target:

$$\mathcal{L}(\mathbf{x}, y, \mathbf{w}_t) = \frac{1}{2}\left(y - \mathbf{w}_t^T\mathbf{x}\right)^2 \tag{2.4}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha \nabla_{\mathbf{w}_t} \mathcal{L}(\mathbf{x}, y, \mathbf{w}_t) \tag{2.5}$$

$$= \mathbf{w}_t + \alpha\left(y - \mathbf{w}_t^T\mathbf{x}\right)\mathbf{x} \tag{2.6}$$

where α is a learning rate parameter.

The Rescorla-Wagner model demonstrated that many empirical observations can be explained through error-based learning. However, there are also important aspects of human learning that the model does not capture (Miller et al., 1995, Daw et al., 2008, Gershman, 2015). Despite constituting an important step in improving our theoretical understanding of human learning, the model does not fully address any of the three principles:

1. It does not guarantee that generalization error is minimal after learning.

2. It can not be adapted to properties of the environment.

3. It does not involve any measure of simplicity.

2.2 Bayesian Inference

From a probabilistic perspective the predictive posterior distribution $p(y|\mathbf{x}, \mathbf{x}_{1:t}, y_{1:t})$ expresses how to make predictions on new inputs given that one has previously observed a sequence of input-target pairs. In this view, building a model of learning boils down to defining how the predictive posterior distribution is obtained.

Bayesian inference offers a principled tool for doing exactly this. Essential for Bayesian inference are a prior $p(\mathbf{w})$ that defines an agent's initial beliefs about possible parameter values before any data is observed and a likelihood $p(y_{1:t}|\mathbf{x}_{1:t}, \mathbf{w})$ that captures its knowledge about how the environment generates data for a given set of parameters. Bayesian inference proceeds to compute the predictive posterior distribution in two steps. First, it combines prior and likelihood to a posterior distribution over parameters by applying Bayes' theorem:

$$p(\mathbf{w}|\mathbf{x}_{1:t}, y_{1:t}) = \frac{p(y_{1:t}|\mathbf{x}_{1:t}, \mathbf{w})p(\mathbf{w})}{p(y_{1:t}|\mathbf{x}_{1:t})} \tag{2.7}$$

Then, it averages over all possible parameter values weighted by their posterior probability to get the predictive posterior distribution:

$$p(y|\mathbf{x}, \mathbf{x}_{1:t}, y_{1:t}) = \int p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathbf{x}_{1:t}, y_{1:t})d\mathbf{w} \tag{2.8}$$

Bayesian theories of human learning are supported by a considerable amount of empirical evidence (Tenenbaum and Griffiths, 2001, Griffiths et al., 2008). They, for example, capture patterns in associative learning that are not accounted for by the Rescorla-Wagner model (Dayan and Long, 1998, Dayan and Kakade, 2001, Courville et al., 2004, 2005, 2006, Gershman, 2015). They have also contributed to our theoretical understanding by unifying theories that have been previously considered as disjunct (Lucas et al., 2015, Anderson, 1991a). Beyond learning, Bayesian inference has been also successfully applied to other domains of human cognition, including perception (Knill and Richards, 1996), motor control (Körding and Wolpert, 2004), everyday judgements (Griffiths and Tenenbaum, 2006) and logical reasoning (Oaksford et al., 2007).

There are, however, also issues that come with Bayesian inference. Applying Bayes' theorem to compute posterior distributions is only possible in a few special cases and uncomputable in general. The main reason for this is that it is often difficult to find a closed-form expression for the integral in the denominator of Equation 2.7. Besides that, it has been argued by critics of the Bayesian approach that characterizing the right prior and likelihood is often impossible. Savage (1972) referred to situations in which this is possible as small worlds and contrasts them with large worlds, in which not all available hypothesis and choices can be enumerated or are known in advance (Gigerenzer and Gaissmaier, 2011). The critics of the Bayesian approach to human learning argue that "real world problems of theoretical significance tend not to be small world problems" (Brighton and Gigerenzer, 2012) and that Bayesian inference provides no justification for optimal reasoning within the real world (Binmore, 2007).

How do Bayesian models address the three suggested principles? If we set our loss to the negative probability of targets under the predictive posterior distribution, Bayesian inference implements an ideal observer, which is optimal in terms of its generalization loss (Jaynes, 2003, Berger, 2013). Bayesian inference can furthermore adapt itself to different environments by adjusting which prior and likelihood it uses. Finally, it is also possible to define priors that favor simple solutions (MacKay, 1992, Rasmussen and Ghahramani, 2001). However, this does not tell us anything about how costly it is to implement the algorithm that computes the solution. Consequently, Bayesian inference only embodies a restricted form of simplicity.

2.3 Rational Process Models

The intractability of Bayesian inference provides a challenge for statisticians, that want to apply such models to real-world problems, and to cognitive scientists, that like to consider them as models of human cognition. In consequence, researchers have developed approaches that can approximate Bayesian inference without running into computational difficulties. The prime examples of such approaches are variational approximations (Jordan et al., 1999) and sample-based methods (Geman and Geman, 1984).

In the context of cognitive science, such approximations are also referred to as *rational process models* (Sanborn et al., 2010, Griffiths et al., 2015). Rational process models can account for cases where human behavior deviates from the notion of optimality prescribed by Bayesian inference. There is evidence that supports both variational and sample-based approximations. For example, variational approximations have been shown to replicate known sensitivities to the arrangement of observations in associative learning studies (Daw et al., 2008, Sanborn and Silva, 2013), whereas sample-based approximation have the potential to account for differences between individual participants (Courville and Daw, 2008, Sanborn et al., 2010, Vul et al., 2014).

Like Bayesian inference, rational process models embody the principles of generalization and adaptation. They are furthermore designed to require fewer computational resources. Therefore, they are simple in some sense. In general, however, rational process models do not comply with the formal definition of simplicity given above – they do not take the length of the computer program that implements the learning algorithm into consideration.

2.4 Program Induction

Learning through *program induction* extends the framework of Bayesian inference to rich, structured hypothesis spaces instead of using simple priors over atomic hypotheses. Potential choices of hypothesis spaces include probabilistic grammars (Goodman et al., 2008), logic systems (Piantadosi et al., 2016) and programming languages (Ellis et al., 2016). The combination of Bayesian inference and structured hypothesis spaces has helped us to understand how people learn and reason in complex domains. Program induction, for example, captures how people learn about compositional concepts from few observations (Lake et al., 2015), and accounts for many patterns in human concept learning (Piantadosi et al., 2016, Rule et al., 2018).

Because program induction is essentially Bayesian inference within a richer space of hypotheses, all statements regarding Bayesian inference and the three principles also apply to program induction. Namely, it does generalize optimally to novel observations and can be adapted to environment-specific characteristics, but only embodies a restricted form of simplicity.

2.5 Connectionism

The framework of *connectionism* (McClelland et al., 1986) provides a very different approach for understanding human learning: it tries to explain learning through neural networks. Typically, these networks consist of a large number of simple processing elements that communicate with each other by transmitting signals. Thus, connectionism explains cognitive processes as emergent consequences of the interaction between a large number of simple processing units, instead of characterizing them directly as it is common in other theories (McClelland et al., 2010).

Neural networks are not a theory of learning on their own; they have to be coupled with a learning algorithm. In the connectionism framework, this is commonly done through gradient-based learning (Rumelhart et al., 1986). Learning in these systems is slow, and as such connectionist models have been traditionally applied to study human learning at larger time-scales; for example, to understand the acquisition of language during development (Elman, 1993, Rumelhart and McClelland, 1986). However, there also exist a number of notable connectionist models in associative learning (Kruschke, 2001), function learning (DeLosh et al., 1997) and category learning (Kruschke, 1992).

Empirically, neural networks show some forms of generalization, but often fail to generalize systematically (Fodor et al., 1988, Lake and Baroni, 2018). More specifically, they do not guarantee that generalization error is minimal after learning. Traditional connectionist models rely on fixed learning algorithms and thus contain no mechanism for adaptation to environment-specific characteristics. They also do typically not involve any measure of simplicity.

2.6 Meta-Learning

The framework of meta-learning offers an alternative to fixed, hand-coded learning algorithms used by connectionist models. The goal of a meta-learning system is to learn the learning algorithm itself through repeated encounters with similar learning problems (Schmidhuber et al., 1996, Bengio et al., 1990, Thrun and Pratt, 1998). The two dominant meta-learning approaches are *gradient-based* *meta-learning* (Finn et al., 2017) and *model-based meta-learning* (Hochreiter et al., 2001b, Santoro et al., 2016).

Model-agnostic meta-learning (MAML) is an example of a gradient-based meta-learning method. MAML attempts to find weight initializations for neural networks that facilitate optimal gradient-based learning within a given environment (Finn et al., 2017, Grant et al., 2018). This results in a connectionist-like learning algorithm that is adapted to environment-specific characteristics. However, while this addresses the issue of slow learning in connectionist models, learning itself is still restricted to gradient-based updates, which means that there is no guarantee that generalization will be optimal after learning. Like traditional connectionist models, MAML also does not involve any measure of simplicity.

Model-based meta-learning is another method to obtain an environment-specific learning algorithm. In this approach, the learning algorithm is represented through a general-purpose function approximator, typically some form of neural network (Hochreiter et al., 2001b, Santoro et al., 2016, Garnelo et al., 2018). The goal is then to turn this function approximator into an optimal learning algorithm. Typically, this is done by minimizing a sample-based estimate of Equation 2.3. If the function approximator is expressive enough, this will lead to an adapted algorithm that generalizes optimally to future observations. Existing model-based meta-learning approaches, however, still do not involve any measure of simplicity.

The study of meta-learning in the context of human learning is still in its infancy, but it has received an increased amount of interest from cognitive science (Griffiths et al., 2019) and neuroscience (Wang et al., 2018) in the recent past.

Whereas both gradient-based and model-based meta-learning have been used to replicate human-like abilities qualitatively (Lake, 2019, McCoy et al., 2020), the work presented in this thesis is among the first to connect meta-learning models to empirical data collected in psychological studies.

2.7 Summary

There have been a number of computational theories of human learning in the past, some of which I have reviewed in this chapter. Each offers its own strengths and weaknesses. Learning algorithms in the Bayesian family generalize optimally to future observations and can be adapted to work well in a given environment. Traditional connectionist models trained through gradient-based methods enjoy no formal guarantees on how well they generalize, although they often show interesting generalization patterns. Meta-learning allows neural network-based models to adapt to an environment through interactions with similar learning problems.

None of the reviewed theories accounts for the cost of implementing the learning algorithm, and thus none of them embodies all three of the suggested principles of human learning. How can one even quantify how costly it is to implement an algorithm? In general, the answer to this question is far from trivial (Li et al., 2008). However, as I will show in the next chapter, there exists a straightforward extension to model-based meta-learning that allows us to do exactly this.

3

A New Modeling Framework

In this chapter, I present two novel classes of learning algorithms; one for solving supervised learning problems called *bounded meta-learned inference* (BMI) and one for solving reinforcement learning problems called RL^3. Both of them embody all three suggested principles of human learning. They are obtained by combining existing model-based meta-learning approaches (Hochreiter et al., 2001a, Santoro et al., 2016, Garnelo et al., 2018, Duan et al., 2016, Wang et al., 2016) with an objective that controls for the description length of the emerging learning algorithm (i.e., the number of bits required to implement it). Section 3.1 introduces general concepts of the framework in the context of supervised learning problems. Section 3.2 demonstrates that the same ideas can also be applied to reinforcement learning problems. Section 3.3 outlines several details that are important for implementing these models.

3.1 Bounded Meta-Learned Inference

Bounded meta-learned inference (BMI) is a class of learning algorithms for supervised learning problems. It combines model-based meta-learning with the minimum description length principle. I will first show how meta-learning can be used to create learning algorithms that generalize optimally in a particular environment. This leads to a variant called *meta-learned inference* (MI). Then, I will show how MI can be extended to the take description length of the emerging learning algorithm into account.

The goal of a learning algorithm is to compute a predictive posterior distributions $p(y|\mathbf{x}, \mathbf{x}_{1:t}, y_{1:t})$. Like Bayesian inference, MI computes statistically optimal predictive posterior distributions, but it does so in a very different way. In MI, a function approximator is trained to act as an optimal learning algorithm – a process that is commonly called meta-learning. How does this work? Initially, the function approximator maps a sequence of previously observed input-target examples $\mathbf{x}_{1:t}, y_{1:t}$ and a queried input $\mathbf{x}$ to a random predictive posterior distribution over targets y. During meta-learning, the system is then trained on a distribution over tasks $p(\mathbf{x}_{1:T}, y_{1:T})$ to infer statistically optimal predictive posterior distributions. I also refer to the distribution over tasks as environment and meta-learning distribution throughout this thesis. In probabilistic terms, we can learn to infer statistically optimal predictive posterior distributions for an environment by minimizing negative log-probabilities of the data under an expectation over tasks:

$$\mathcal{L}_{\text{MI}}(\mathbf{\Theta}) = \mathbb{E}_{p(\mathbf{x}_{1:T}, y_{1:T})}\left[\sum_{t=0}^{T-1} - \log p(y_{t+1} | \mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \mathbf{\Theta})\right] \tag{3.1}$$

where $\mathbf{\Theta}$ denotes the parameters of the function approximator. In principle, any type of general-purpose function approximator could be used to implement this mapping. However, as the length of input-target sequences may vary, *recurrent neural networks* (RNNs) are a natural choice for the aforementioned sequential learning setting. Figure 3.1 shows an example of how the RNN processes a sequence of data-points.

During meta-learning, a sample-based approximation of the MI objective (Equation 3.1) is optimized until convergence using standard optimization techniques. Figure 3.2 provides pseudocode describing the meta-learning procedure. Through repeated exposure to tasks from the meta-learning distribution, the model adapts itself to the properties of the encountered environment. After meta-learning is

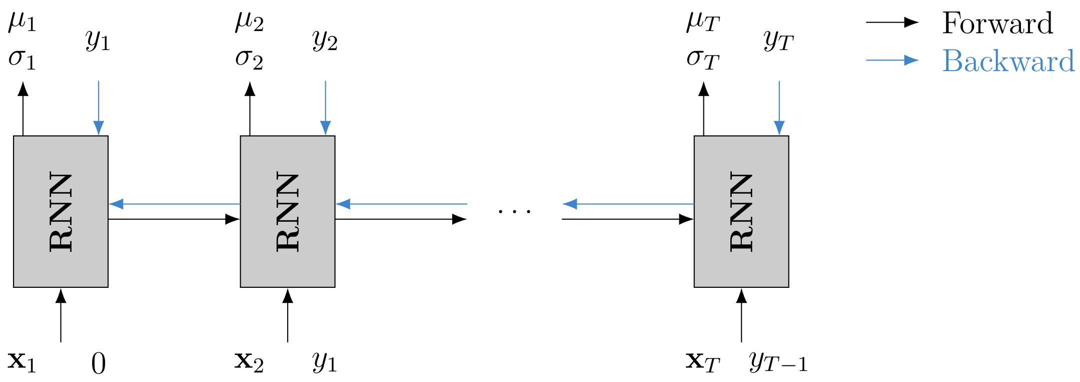


Figure 3.1: Unrolled RNN processing a single task $\mathbf{x}_{1:T}, y_{1:T}$. In time-step $t+1$ the network reads in the input $\mathbf{x}_{t+1}$ and the target from the previous time-step y_t. The outputs μ_{t+1}, σ_{t+1} correspond to parameters of the predictive posterior distribution $p(y_{t+1}|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \mathbf{\Theta}) = \mathcal{N}(y_{t+1}|\mu_{t+1}, \sigma_{t+1})$. For classification tasks the predictive posterior distribution may be parametrized through a categorical distribution. Black arrows indicate forward passes, blue arrows indicate backward passes.

completed, the RNN acts as a free-standing learning algorithm without requiring any further parameter updates. Instead, learning is implemented through the forward dynamics of the RNN: we provide the network with a sequence of input-target examples and an input that we want to query, and the network provides us with predictions for that input. Typically, one is only interested in the properties of the emerging learning algorithm after meta-learning is completed and not in what happens during meta-learning. Thus, all results reported in this thesis refer to fully converged models.

MI is a learning algorithm that embodies two of the suggested principles of human learning; generalization and adaptation. The model learns to generalize optimally because the log-probability term inside the expectation encourages it to make optimal predictions on future observations. The model is adapted to environment-specific characteristics because this term is optimized for a particular distribution over tasks. It has been shown in previous work that this

```
Algorithm 1: Meta-Learning
while not converged do
    /* sample a batch of tasks */
    $\mathbf{x}_{1:T}, y_{1:T} \sim p(\mathbf{x}_{1:T}, y_{1:T})$;
    /* compute loss */
    $\mathcal{L}_{\text{MI}}(\boldsymbol{\Theta}) = \sum_{t=0}^{T-1} - \log p(y_{t+1} | \mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \boldsymbol{\Theta})$;
    /* update model parameters */
    $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \alpha \nabla_{\boldsymbol{\Theta}} \mathcal{L}_{\text{MI}}(\boldsymbol{\Theta})$;
end
```

Figure 3.2: Pseudocode for the meta-learning procedure. For BMI, one has to replace the MI objective with the BMI objective.

meta-learning approach leads to the emergence of a learning algorithm that approximately simulates Bayesian inference (Rabinowitz, 2019, Ortega et al., 2019, Mikulik et al., 2020). I provide several model simulation results that support this claim in Section 3.1.1.

BMI combines the MI objective with an additional term that acts as a regularizer for the description length of the emerging learning algorithm. Therefore, it also embodies the third suggested principle of human learning; simplicity. In the meta-learning setting, the trained RNN acts as a learning algorithm. Thus, controlling the description length of RNN parameters implies that we control how many bits are required to implement the learning algorithm itself.

How can we accomplish this? Instead of keeping point estimates of neural network parameters, we keep track of a distribution over their plausible values $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$, which I refer to as the encoding distribution. During meta-learning, we then attempt to optimize the parameters of this encoding distribution $\boldsymbol{\Lambda}$.[*] For

[*] For example, $\boldsymbol{\Lambda}$ could correspond to the mean and standard deviation of a normal distribution.

BMI, this involves minimizing the following objective:

$$\mathcal{L}_{\text{BMI}}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\mathcal{L}_{\text{MI}}(\boldsymbol{\Theta})\right] + \beta \text{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})||p(\boldsymbol{\Theta})\right] \tag{3.2}$$

where $p(\boldsymbol{\Theta})$ is a prior over model parameters. The *Kullback–Leibler* (KL) divergence between $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$ and $p(\boldsymbol{\Theta})$ can be interpreted as a measure of the parameters' description length (Hinton and Van Camp, 1993).[†] Minimizing the BMI objective (Equation 3.2), therefore, leads to a learning algorithm that optimally trades-off predictive power against β-weighted description length. For $\beta = 0$, it reverts to an unconstrained learning algorithm that is equivalent to MI.

It is possible to map the three principles neatly onto different parts of the BMI objective:

$$\mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\underbrace{\mathbb{E}_{p(\mathbf{x}_{1:T}, y_{1:T})}}_{\text{adaptation}}\left[\underbrace{\sum_{t=0}^{T-1} -\log p(y_{t+1}|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \boldsymbol{\Theta})}_{\text{generalization}}\right]\right] + \underbrace{\beta\text{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})||p(\boldsymbol{\Theta})\right]}_{\text{simplicity}} \tag{3.3}$$

BMI achieves generalization by being explicitly trained to make optimal inferences on new data-points after observing a sequence of input-target examples from the same task. BMI achieves adaptation by maximizing performance for a specific distribution over tasks. BMI achieves simplicity by penalizing the description length of the emerging learning algorithm.

For completeness, it should be noted that there also exists an interpretation of the BMI objective that appeals to *PAC-Bayesian theory* (McAllester, 1999, 2013,

[†] Further explanation about why that is the case is provided Section 3.3.2.

Dziugaite and Roy, 2017, Achille and Soatto, 2018). From the PAC-Bayesian perspective, minimizing the BMI objective corresponds to minimizing an upper-bound on the generalization error for tasks that were not encountered during meta-learning. Recently, several meta-learning models have been proposed that that rely on this alternative interpretation instead of the information-theoretic interpretation used here (Yin et al., 2019, Rothfuss et al., 2020, Jose and Simeone, 2020).

3.1.1 Meta-Learning and Bayes-Optimality

Previously, I have remarked that MI – or equivalently BMI without resource limitations – approximately simulates a Bayes-optimal learner. This statement is of course subject to the expressiveness of the employed function approximator, and one might question whether it also holds in practice. In this section, I demonstrate that this is indeed the case on a simple example problem. Let us consider the following environment:

$$p(\mathbf{x}_{1:T}, y_{1:T}, \mathbf{w}) = p(\mathbf{w}) \prod_{t=1}^{T} p(\mathbf{x}_t) p(y_t|\mathbf{x}_t, \mathbf{w}) \tag{3.4}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) \tag{3.5}$$

$$p(\mathbf{x}_t) = \mathcal{U}(\mathbf{x}_t| - 1, 1) \tag{3.6}$$

$$p(y_t|\mathbf{x}_t, \mathbf{w}) = \mathcal{N}(y_t|\mathbf{w}^T\mathbf{x}_t, 0.1) \tag{3.7}$$

Equations 3.4 to 3.7 state that tasks in this environment are generated by first sampling a random weight vector, then sampling random input vectors $\mathbf{x}_{1:T}$, and finally sampling their corresponding targets $y_{1:T}$. They also express the assumptions that the input-target relationship is linear and that each input-target pair is independent of other input-target pairs given the weight vector.

An agent sequentially observes input-target examples and has to make predictions for new inputs without having access to the underlying weight vector. A Bayes-optimal learner does this by computing the predictive posterior distribution. The given environment is simple enough such that it is possible to find an analytical expression for the predictive posterior distribution (Dayan and Kakade, 2001, Gershman, 2015):

$$p(y_t|\mathbf{x}_t, \mathbf{x}_{1:t-1}, y_{1:t-1}) = \int p(y_t|\mathbf{x}_t, \mathbf{w})p(\mathbf{w}|\mathbf{x}_{1:t-1}, y_{1:t-1})d\mathbf{w} \tag{3.8}$$

$$= \mathcal{N}(y_t|\boldsymbol{\mu}_t^T\mathbf{x}_t, \mathbf{x}_t^T\boldsymbol{\Sigma}_t\mathbf{x}_t + \sigma^2) \tag{3.9}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{k}_t\left(y_{t-1} - \boldsymbol{\mu}_{t-1}^T\mathbf{x}_{t-1}\right) \tag{3.10}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1} - \mathbf{k}_t\mathbf{x}_{t-1}^T\boldsymbol{\Sigma}_{t-1} \tag{3.11}$$

$$\mathbf{k}_t = \frac{\boldsymbol{\Sigma}_{t-1}\mathbf{x}_{t-1}}{\mathbf{x}_{t-1}^T\boldsymbol{\Sigma}_{t-1}\mathbf{x}_{t-1} + \sigma^2} \tag{3.12}$$

Next, I will show that MI closely approximates this Bayes-optimal learner. The implementation used here makes use of *gated recurrent units* (Bahdanau et al., 2014), which are a particular type of RNN.[‡] The network processes the current input $\mathbf{x}_t$ and the last target y_{t-1}, and outputs the parameters of a normal distribution over y_t as shown in Figure 3.1. In the ideal case, these parameters should be identical to the ones of the predictive posterior distribution.

The sequence length was fixed to $T = 10$ for all simulations, and I report results for environments of varying complexity. The complexity of an environment

[‡] For further details on gated recurrent units, see Section 3.3.1.

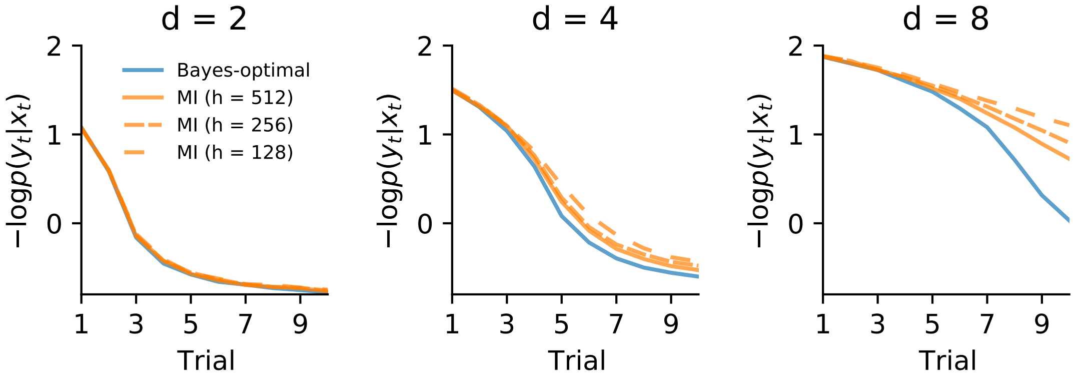


Figure 3.3: Performance comparison of MI with the Bayes-optimal learner in environments of different complexity. The x-axis indicates the number of observed data-points, the y-axis negative log-probabilities of targets under the predictive distribution averaged over 10000 tasks.

is controlled by the dimensionality of the weight vector $d \in \{2, 4, 8\}$. The MI objective 3.1 was minimized during meta-learning using the AmsGrad optimizer (Reddi et al., 2019) with a learning rate of 0.001. In total, meta-learning comprised $2.5 \cdot 10^6$ gradient steps with batches of size 32.

Figure 3.3 compares the performance of MI with $h \in \{128, 256, 512\}$ hidden units to the Bayes-optimal learner. For the environment with the lowest complexity ($d = 2$), the performance of MI is identical to the Bayes-optimal learner for all practical purposes. For the environment with medium complexity ($d = 4$), the performance of MI deviates only minimally from the Bayes-optimal learner. For the environment with the highest complexity ($d = 8$), the performance gap becomes larger; at this point MI does not anymore align with the Bayes-optimal learner. The studies presented later on involve between one and four features. It can therefore be concluded that MI – or equivalently BMI without resource limitations – approximately simulates a Bayes-optimal learner for the purpose of this thesis.

3.1.2 Why Not Bayesian Inference?

It is reasonable to ask what additional explanatory power the proposed framework offers, given that meta-learning produces models that approximately simulate Bayesian inference. After all, why not just stick to good-old Bayesian inference? There are, however, a few arguments that speak in favor of BMI.

First, BMI represents the learning algorithm with a parametric model, which makes it possible to quantify – and control for – its complexity in a principled way. It can, therefore, *also* explain behavior that deviates from the notion of optimality prescribed by Bayesian inference (Kahneman and Tversky, 1972, Chater and Vitányi, 2003, Gershman et al., 2015).

Second, Bayesian inference scales poorly to complex situations to the extent that finding analytical expressions for predictive posterior distributions is not possible beyond idealized toy examples. BMI, on the other hand, can be trained to infer approximately optimal predictive posterior distributions, even if the corresponding Bayesian inference problem has no analytical solution. It also makes the design of Bayes-optimal learning algorithms easier and faster, as one is not forced to derive new updating equations anymore whenever encountering a new inference problem. Instead, one can run the same meta-learning algorithm in a new environment to automate the process of creating a Bayes-optimal learning algorithm.

Third, Bayesian inference requires access to an explicit functional form of the likelihood and the prior. If these are not available, Bayesian inference provides no justification for optimal reasoning (Binmore, 2007). BMI, on the other hand, can infer approximately optimal predictive posterior distributions even if it is impossible to phrase the corresponding inference problem in the first place; all that it needs for this are samples of tasks from the environment. In contrast to Bayesian inference, BMI can therefore provide a justification for optimal reasoning in large world problems.

3.1.3 Marr's Levels of Analysis

Nearly 40 years ago, Marr (1982) presented a framework for categorizing modeling approaches depending on their level of analysis. He suggested three such levels: computational, algorithmic, and implementational. The computational level is about defining what goal the system under investigation is trying to achieve. The algorithmic level describes how the system transforms inputs into outputs, or in other words, how it processes information. Finally, we have the implementational level that specifies how the system is realized in physical hardware. Not all modeling approaches map nicely on Marr's levels of analysis, but many do. Indeed, up to this date, Marr's levels continue to be an influential framework in cognitive science for thinking about what questions to ask and how to answer them (Griffiths et al., 2012, Zednik and Jäkel, 2016, Lieder and Griffiths, 2020).

Marr's levels help us to clarify what questions a given model tries to answer. Large parts of contemporary cognitive science are devoted to identifying the processes that describe how the brain maps inputs into outputs. The resulting process models tend to be located at Marr's algorithmic level (Jarecki et al., 2020). For example, Lewandowsky and Farrell (2010) argued that "cognitive scientists would ultimately prefer an explanatory process model over most characterizations", and McClelland (2009) stated that "most cognitive scientists are concerned with understanding cognitive processes". If one is looking for such a model, BMI will probably not provide a satisfying answer. Instead, BMI answers a question on Marr's computational level of analysis: what would an optimal learner that is subject to limited computational resources do in a particular environment?

3.2 RL^3

The ideas that were outlined in the last section can not only be applied to supervised learning problems but also to reinforcement learning problems (Sutton and Barto, 2018). Reinforcement learning deals with sequential decision-making problems in unknown environments, which are typically modeled as *Markov decision processes* (MDPs, Bellman, 1957). Let $M = (\mathcal{S}, \mathcal{A}, p)$ be an undiscounted MDP with a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$ and a joint distribution over the next state and a scalar reward signal $p(s_{t+1}, r_t | s_t, a_t)$. Note that this formulation does not include a discount factor γ. However, discounting can always be incorporated if desired by modifying the transition dynamics such that the agent transitions to a terminal state with probability $1 - \gamma$ (Levine, 2018).

A reinforcement learning agent interacts with an MDP in discrete time intervals. In each time-step t, it observes the current state of environment s_t, based on which it executes a specific action a_t. In turn, this action influences the state of the environment and triggers a reward signal r_t. The interaction between agent and environment repeats for a certain number of time-steps or until a terminal state is reached. The agent's goal is to find the policy π that maximizes the expected return $\mathbb{E}_{p,\pi}\left[\sum_{t=1}^{\infty} r_t\right]$ without having access to the dynamics of the environment $p(s_{t+1}, r_t | s_t, a_t)$.

How can one meta-learn an algorithm for solving MDPs? Previous work has shown that it is possible to accomplish this through standard reinforcement learning techniques (Wang et al., 2016, Duan et al., 2016). Duan et al. (2016) refer to this approach as RL^2. In this section, I will present a resource-limited extension of RL^2 to which I refer to as RL^3. RL^3 shares many of its ideas with BMI. As in BMI, a RNN is trained on a distribution over tasks to act as a free-standing learning algorithm, but this time to solve an MDP sampled from a distribution over MDPs instead of solving a supervised learning problem. The RNN takes previous actions and rewards as inputs in addition to the current state, making the output a function of the current state s_t and the entire history $h_t = s_{1:t-1}, a_{1:t-1}, r_{1:t-1}$. By integrating information from the history, the RNN attempts to infer the optimal policy for the currently encountered MDP.

Any reinforcement learning algorithm could be used to turn the RNN into an optimal reinforcement learning algorithm. Throughout this thesis, I make use of a particular algorithm called *Q-Learning* (Watkins and Dayan, 1992). The goal of Q-Learning is to learn the action-value function of the optimal policy π^*:

$$Q^*(s_t, h_t, a_t) = \mathbb{E}_{p,\pi^*}\left[\sum_{k=0}^{\infty} r_{t+k} | s_t, h_t, a_t\right] \tag{3.13}$$

If one has access to Q^*, the optimal policy can be readily derived:

$$\pi^*(a_t|s_t, h_t) = \begin{cases} 1 & \text{if } a_t = \arg\max_{a \in \mathcal{A}} Q^*(s_t, h_t, a) \\ 0 & \text{else} \end{cases} \tag{3.14}$$

Keeping the probabilistic interpretation from the earlier section, the outputs of the RNN parametrize a distribution over action-values of the optimal policy, and model parameters are adjusted to maximize the log-likelihood of observed data-points. However, one cannot use Q^* directly as targets because that would require access to the optimal policy. The key insight behind Q-Learning is that one can instead use targets defined by:

$$\begin{aligned} q_t &= r_t + \max_a Q^*(s_{t+1}, h_{t+1}, a) && (3.15) \\ &\approx r_t + \max_a \mathbb{E}\left[q_{t+1} | s_{t+1}, h_{t+1}, a, \boldsymbol{\Theta}\right] && (3.16) \end{aligned}$$

For tabular environments, it can be shown that using targets as defined in Equation 3.15 leads to provable convergence to the action-value function of the optimal policy (Jaakkola et al., 1994). Q^* can be estimated by a simple forward pass through the network after meta-learning is completed, and one can subsequently derive the optimal policy based on Equation 3.14. Therefore, the RNN now implements a freestanding reinforcement learning algorithm through its recurrent activations.

The complete RL^3 objective is obtained by trading-off accurate predictions of optimal action-values for a shorter description length:

$$\begin{aligned} \mathcal{L}_{\mathrm{RL}^3}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})} & \left[\mathbb{E}_{p(s_{1:T}, a_{1:T}, r_{1:T})} \left[\sum_{t=0}^{T-1} - \log p(q_{t+1} | s_{t+1}, h_{t+1}, a_{t+1}, \boldsymbol{\Theta}) \right] \right] \\ & + \beta \mathrm{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}) || p(\boldsymbol{\Theta}) \right] \qquad (3.17) \end{aligned}$$

Minimizing the RL^3 objective for $\beta = 0$ leads to a reinforcement learning algorithm that approximates the Bayes-optimal policy (Duff, 2003, Zintgraf et al., 2019b, Ortega et al., 2019). For $\beta > 0$, we get a family of reinforcement learning algorithms that trade-off performance for a shorter description length. Like its counterpart for solving supervised learning problems, RL^3 also embodies all three suggested principles of human learning; generalization, adaptation, and simplicity.

3.3 Modeling Details

Thus far, I have introduced the general ideas of BMI and RL^3, but remained vague about some of their details. In this section, I will close this gap. Each subsection focuses on a particular aspect of the BMI or RL^3 objective. Section 3.3.1 provides details about the model architecture that is used. Section 3.3.2 explains why the KL divergence between encoding distribution and prior can be interpreted as a measure of the parameters' description length. Section 3.3.3 specifies the encoding distribution and prior used. Finally, Section 3.3.4 provides a short discussion about how to select the appropriate meta-learning distribution when considering BMI and RL^3 as models of human learning.

3.3.1 Gated Recurrent Units

I have argued that RNNs are the natural choice for a function approximator in our setting as they can easily deal with sequences of varying lengths. This is important because it allows us to condition the model on a variable number of previously observed input-target examples. RNNs do this by compressing a history of observations of arbitrary length into a vector of fixed length $\mathbf{h}_t \in \mathbb{R}^h$. In each time-step, this hidden state is updated based on the new inputs to the model

$\mathbf{x}_t, y_{t-1}$ and the hidden state from the previous time-step $\mathbf{h}_{t-1}$. Early RNNs were based on simple activation functions (Elman, 1990), such as the tanh function:

$$\mathbf{h}_t = \tanh\left(\mathbf{W}_{ih}[\mathbf{x}_t, y_{t-1}] + \mathbf{W}_{hh}\mathbf{h}_{t-1}\right) \tag{3.18}$$

where $\mathbf{W}_{ih}$ denotes parameters of a linear transformation from the input to the hidden state and $\mathbf{W}_{hh}$ denotes parameters of a linear transformation from the previous hidden state to the updated hidden state.

RNNs are typically trained using gradient-based methods. The most popular approach to obtain gradients with respect to RNN parameters is the idea of *backpropagation through time* (Werbos, 1988). In backpropagation through time, the RNN is unrolled over multiple time-steps, which results in a computational graph that can be interpreted as a very deep feed-forward network with shared weights. In principle, this allows to directly compute the desired gradients using automatic differentiation software. However, in practice this type of training suffers from the problem of vashining gradients (Hochreiter et al., 2001a); gradients of the loss function will approach zero as the depth of the network increases, essentially preventing the model from learning long-term dependencies.

A popular approach to prevent this issue is to use gated activations functions, which better control the information flow through the hidden state (Hochreiter and Schmidhuber, 1997, Chung et al., 2014). Among RNNs with gated activations functions the *gated recurrent unit* (GRU, Bahdanau et al., 2014) has become one of the standard models. It is based on the following updating equation for the hidden state:

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_{ir}[\mathbf{x}_t, y_{t-1}] + \mathbf{W}_{hr}\mathbf{h}_{t-1}\right) \tag{3.19}$$

$$\mathbf{z}_t = \sigma\left(\mathbf{W}_{iz}[\mathbf{x}_t, y_{t-1}] + \mathbf{W}_{hz}\mathbf{h}_{t-1}\right) \tag{3.20}$$

$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \tanh\left(\mathbf{W}_{ih}[\mathbf{x}_t, y_{t-1}] + \mathbf{W}_{hh}\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right) \tag{3.21}$$

where σ denotes the logistic sigmoid function, $\odot$ denotes element-wise multiplication and $\boldsymbol{\Theta} = \{\mathbf{W}_{ir}, \mathbf{W}_{hr}, \mathbf{W}_{iz}, \mathbf{W}_{hz}, \mathbf{W}_{ih}, \mathbf{W}_{hh}\}$ are model parameters.

All meta-learning models used in this thesis are based on GRUs. GRUs are a suitable choice because they are Turing-complete (Siegelmann and Sontag, 1992), and because they have been shown to work well in a diverse set of problems. However, there exists a range of alternatives that could also be successfully applied within the same setting, such as models with external memory (Graves et al., 2016) or attention-based models (Mishra et al., 2017, Kim et al., 2019).

3.3.2 The Bits-Back Argument

Why does minimizing the KL divergence lead to models with a shorter description length? The formal justification for this is given by the *bits-back argument* (Hinton and Van Camp, 1993). The goal of this section is to explain how bits-back coding schemes compress data. To understand bits-back coding, however, it is useful to discuss simpler coding schemes first. Thus, I will explain both deterministic and stochastic two-part coding schemes before getting to the bits-back argument.

Let me start by assuming that we would like to find the shortest description of the data possible; this is known as the *minimum description length* principle (Rissanen, 1978, Hinton and Van Camp, 1993, Grünwald and Grunwald, 2007). Why is that a desirable goal? The thought here is that any regularity in the data can be used to compress it. This implies that if we compress the data enough, we will find regularities in it. Throughout this section, I assume that we are given some data $y \sim p(y)$ that we would like to compress. However, note that all ideas presented here also apply to the conditional probability distributions used in meta-learning. According to Shannon's source coding theorem (Shannon, 1948) transmitting the data losslessly with an optimal code requires on average at least $H\left[p(y)\right] = \mathbb{E}_{p(y)}\left[-\log_2 p(y)\right]$ bits. Efficient coding schemes that come close to this lower bound exist if $p(y)$ is known (Huffman, 1952, Witten et al., 1987).

Deterministic Two-Part Coding Schemes

However, often we do not have access to $p(y)$; learning it is the actual problem. What can be done instead in that case? Two-part coding schemes offer one solution to this problem (Grünwald and Grunwald, 2007). They assume that both encoder and decoder agree in advance on a class of parametrized models $p(y|\boldsymbol{\Theta})$ and a prior over their parameters $p(\boldsymbol{\Theta})$. In a two-part coding scheme, the encoder splits transmitting the data into two parts. First, it selects an arbitrary $\boldsymbol{\Theta}$ and compresses it based on the prior distribution. Then, being aware that the decoder can recover these parameters, it does not have to compress the entire data anymore, but only the error that is not explained by the model with the chosen parameters. This leads to an expected code length of:

$$\mathcal{L}_{\text{det}}(\boldsymbol{\Theta}) = \underbrace{\mathbb{E}_{p(y)}[-\log_2 p(y|\boldsymbol{\Theta})]}_{\substack{\text{bits to transmit}\\ \text{model mismatch}}} \underbrace{-\log_2 p(\boldsymbol{\Theta})}_{\substack{\text{bits to transmit}\\ \text{model parameters}}} \tag{3.22}$$

Note, that minimizing a sample-based estimate of Equation 3.22 is equivalent to finding the maximum a posteriori estimate of model parameters (Honkela and Valpola, 2004).

Stochastic Two-Part Coding Schemes

Alternatively, the encoder might not use a single deterministic $\boldsymbol{\Theta}$, but instead encode parameters based on a sample from an encoding distribution $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$. Doing so leads to the following expected code length:

$$\mathcal{L}_{\text{sto}}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\mathbb{E}_{p(y)}\left[-\log_2 p(y|\boldsymbol{\Theta})\right] - \log_2 p(\boldsymbol{\Theta})\right] \tag{3.23}$$

$$= \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\mathbb{E}_{p(y)}\left[-\log_2 p(y|\boldsymbol{\Theta})\right]\right] + H\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}); p(\boldsymbol{\Theta})\right] \tag{3.24}$$

It holds that $\min_{\boldsymbol{\Lambda}} \mathcal{L}_{\text{sto}}(\boldsymbol{\Lambda}) \geq \min_{\boldsymbol{\Theta}} \mathcal{L}_{\text{det}}(\boldsymbol{\Theta})$. Hence, at first sight it seems like nothing can be gained by using a stochastic parameter encoding.

Bits-Back Coding Schemes

This is where the bits-back argument comes in. It turns out that using a stochastic parameter encoding it is possible to achieve a shorter code length if we make two additional assumptions:

- We want to transmit some additional auxiliary data in the form of random bits.

- The decoder is able to run the same algorithm as the encoder to obtain $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$.

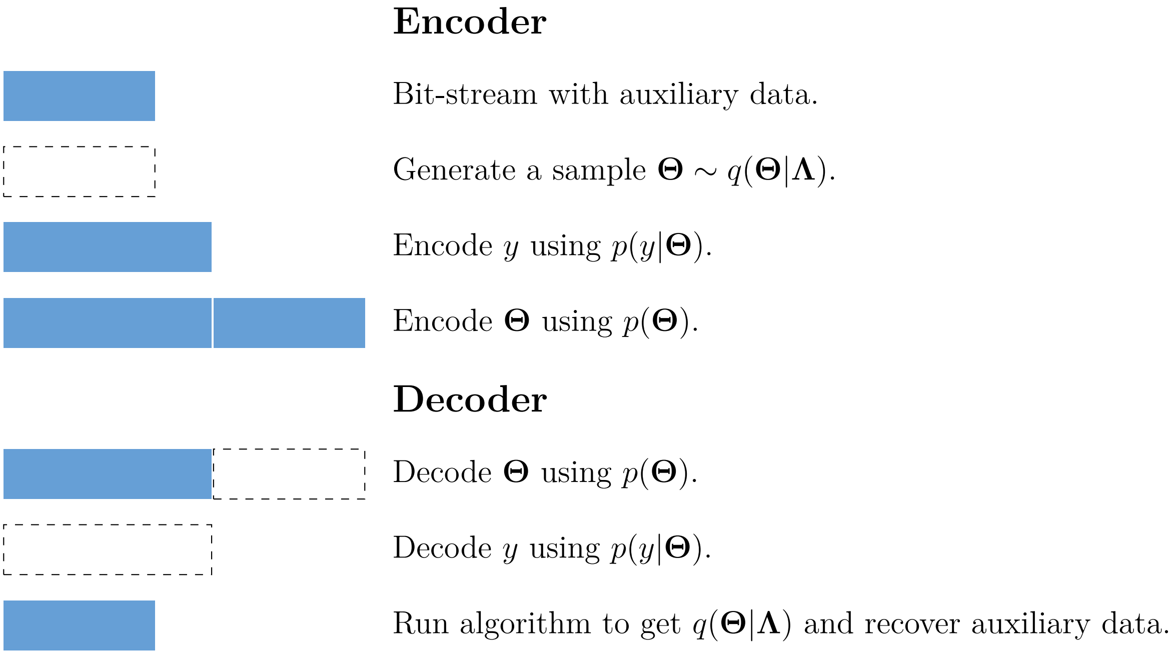
The key observation of bits-back coding is that sampling from a probability distribution is technically equivalent to generating a sequence of random bits. In a bits-back coding scheme, the encoder proceeds in the following steps:

1. Use random bits from the auxiliary data to generate a sample $\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$, which requires to use $-\log_2 q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$ bits from the auxiliary data.

2. Encode the mismatch between model predictions and the data, which leads to a sequence of $-\log_2 p(y|\boldsymbol{\Theta})$ bits.

3. Encode the parameter sample using the prior distribution, which leads to a sequence of $-\log_2 p(\boldsymbol{\Theta})$ bits.

Then, the encoder transmits the entire sequence of bits to the decoder, i.e., the bits for the sampled parameter value and the bits for the modeling mismatch. Importantly, it does not have to transmit the bits from the auxiliary data that were used to generate $\boldsymbol{\Theta}$ because the decoder can recover them by itself. How can it do that? After decoding the transmitted parameter value and data, the decoder can run the same algorithm as the encoder to get $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$. Then, it has access to $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$ and the sample $\boldsymbol{\Theta}$ that was transmitted by the encoder, which in turn allows the decoder to deduce which sequence of bits was used to generate the sample. This process is illustrated graphically in Figure 3.4.



Encoder

Bit-stream with auxiliary data.

Generate a sample $\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$.

Encode y using $p(y|\boldsymbol{\Theta})$.

Encode $\boldsymbol{\Theta}$ using $p(\boldsymbol{\Theta})$.

Decoder

Decode $\boldsymbol{\Theta}$ using $p(\boldsymbol{\Theta})$.

Decode y using $p(y|\boldsymbol{\Theta})$.

Run algorithm to get $q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})$ and recover auxiliary data.

Figure 3.4: Schematic illustration of how the data is encoded and decoded in a bits-back coding scheme. Figure adapted from (Kingma et al., 2019, Townsend et al., 2019).

In expectation, this leads to the following description length of the data:

$$\mathcal{L}_{\text{bb}}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\mathbb{E}_{p(y)}\left[-\log_2 p(y|\boldsymbol{\Theta})\right]] + H\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}); p(\boldsymbol{\Theta})\right] - H\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})\right] \quad (3.25)$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}[\mathbb{E}_{p(y)}\left[-\log_2 p(y|\boldsymbol{\Theta})\right]]}_{\substack{\text{bits to transmit}\\ \text{model mismatch}}} + \underbrace{\text{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})||p(\boldsymbol{\Theta})\right]}_{\substack{\text{bits to transmit}\\ \text{model parameters}}} \quad (3.26)$$

The bits-back argument makes it possible to identify the KL divergence between encoding distribution and prior with the description length of model parameters. It also demonstrates that it can be advantageous to use a stochastic encoding distribution instead of encoding with deterministic parameters. There is a nice intuitive explanation for this phenomenon: for some parameters, knowing their exact value is not so important, and hence we can represent them with limited precision, and a stochastic encoding allows us to do exactly this through the uncertainty represented in the encoding distribution $q(\mathbf{\Theta}|\mathbf{\Lambda})$. Note the immediate similarity between Equation 3.26 and the BMI objective from Equation 3.2. The main difference between both is that the BMI objective adds an additional hyperparameter β for controlling how important compression of parameters is relative to model performance.

3.3.3 Variational Dropout Prior

The bits-back argument holds for any prior, as long it is known by both the encoder and the decoder. However, in general, priors that promote simple functions are favored. The models presented in this thesis use sparsity-inducing priors (Tipping, 2001). This type of prior is appealing for two reasons:

1. They have been successfully applied to a wide range of neural network architectures (Nowlan and Hinton, 1992, Williams, 1995, Blundell et al., 2015, Ghosh and Doshi-Velez, 2017, Louizos et al., 2017, Ullrich et al., 2017).

2. There is evidence from the neuroscience literature that sparsity plays an important role in the human brain (Olshausen and Field, 1997, 2004).

In particular, throughout this thesis, I employ a prior that is based on the idea of *variational dropout* (Kingma et al., 2015). In variational dropout parameter means $\boldsymbol{\mu}_i$ are corrupted by multiplicative normally distributed noise $\boldsymbol{\xi}_i$ as follows:

$$\boldsymbol{\Theta}_i = \boldsymbol{\mu}_i \cdot \boldsymbol{\xi}_i \tag{3.27}$$

$$\boldsymbol{\xi}_i \sim \mathcal{N}(\boldsymbol{\xi}_i|1, \boldsymbol{\alpha}_i) \tag{3.28}$$

$$\Rightarrow q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}) = \prod_i \mathcal{N}(\boldsymbol{\Theta}_i|\boldsymbol{\mu}_i, \boldsymbol{\alpha}_i\boldsymbol{\mu}_i^2) \tag{3.29}$$

Instead of parametrizing the encoding distribution by $\boldsymbol{\Lambda} = \{\boldsymbol{\mu}_i, \boldsymbol{\alpha}_i\}$, Molchanov et al. (2017) suggested the following reparametrization to reduce the variance of stochastic gradients:

$$\boldsymbol{\sigma}_i^2 = \boldsymbol{\alpha}_i \cdot \boldsymbol{\mu}_i^2 \tag{3.30}$$

$$\Rightarrow q(\boldsymbol{\Theta}|\boldsymbol{\Lambda}) = \prod_i \mathcal{N}(\boldsymbol{\Theta}_i|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2) \tag{3.31}$$

which is used together with an improper log-scale uniform prior over model parameters. There is no analytical expression for the KL term (ref. Equation 3.2) under this prior and encoding distribution; however it can be approximated numerically. Molchanov et al. (2017) suggested the following approximation:

$$\mathrm{KL}\left[q(\boldsymbol{\Theta}_i|\boldsymbol{\Lambda}_i)||p(\boldsymbol{\Theta}_i)\right] \approx -k_1\sigma(k_2 + k_3 \log \boldsymbol{\alpha}_i) + 0.5 \log(1 + \boldsymbol{\alpha}_i^{-1}) - \text{const.} \tag{3.32}$$

$$k_1 = 0.63576, k_2 = 1.87320, k_3 = 1.48695 \tag{3.33}$$

How does this prior work? The KL term is minimized as $\boldsymbol{\alpha}_i$ goes towards infinity, thus regularization favors large $\boldsymbol{\alpha}_i$ (Molchanov et al., 2017). Looking at Equation 3.27, one can see that this leads to parameters that are corrupted by large multiplicative noise, which negatively affects performance. The only way to prevent a decrease in performance is to turn off unimportant parameters entirely by setting some subset of $\boldsymbol{\mu}_i$ to zero. In our setting, this implies that under large resource limitations only networks with few non-zero parameters remain. Thus, resulting learning algorithms are not only simple in terms of their description length but also in terms of the number of their remaining parameters.

3.3.4 What Tasks To Adapt To?

Properties of BMI and RL^3 are determined by the distribution over tasks that was used during meta-learning. When considering them as models of how people learn, an important question arises: how to choose the "right" distribution over tasks? For the upcoming studies, I use the following guidelines to construct meta-learning distributions:

1. The meta-learning distribution should reflect people's experiences with similar problems (either through direct interactions or evolutionary processes).

2. The meta-learning distribution should reflect people's expectancies about what tasks they might encounter in the study.

Note that the distribution on which BMI and RL^3 are eventually evaluated might differ from the one used during meta-learning.

4

Heuristics From Bounded Meta-Learned Inference

Abstract: How do people make decisions? An influential theory of decision-making argues that people do so using heuristics – simple strategies that ignore part of the relevant information. We show that BMI discovers two previously suggested types of heuristics – one reason decision-making and equal weighting – in specific environments, and thus provide a normative justification for heuristic decision-making. Furthermore, BMI makes clear and precise predictions about when each heuristic should be applied, which allows us to gain new insights on the mixed results of prior empirical work on heuristic decision-making. In three empirical paired comparison studies with continuous features, we verify predictions of our theory and show that it captures several characteristics of human decision-making not explained by alternative theories.

This chapter is based on the following publication:

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2020). Heuristics From Bounded Meta-Learned Inference (under review).

The code to reproduce the presented results is publicly available under https://github.com/marcelbinz/HeuristicsFromBMLI.

4.1 Introduction

Imagine having to decide which of two movies you are going to watch tonight: Movie A vs. Movie B. Movie A has a higher average rating on a website that you trust, while Movie B is directed by a known director and has previously won an Oscar for the best picture. From past experiences, you know that rating is the best indicator for a good movie. Whether the movie won an Oscar and who directed it is less important for how much you normally enjoy watching a movie. How do people make decisions like this? The question of how people decide between two options is as fundamental as its answer is contentious. Indeed, even though we make countless such decisions every day, the underlying principles of these decisions are still debated in psychology (Todd and Gigerenzer, 2000), behavioral economics (Samuels et al., 2012), and neuroscience (Camerer et al., 2005).

Traditionally, researchers have approached this problem by looking at how rational agents decide. From this ideal observer perspective (Geisler, 1989) it is assumed that people weight different attributes of each option appropriately to combine information from all available sources. Psychologists were however quick to point out that rational decision-making can be too burdensome (Simon, 1990b, Tversky and Kahneman, 1974). Instead, they suggested that human decision-making may be based on a variety of heuristics, which are simple strategies that ignore part of the relevant information (Gigerenzer and Todd, 1999, Shah and Oppenheimer, 2008, Tversky and Kahneman, 1974).

Two common classes of heuristics are *one reason decision-making* (Gigerenzer and Goldstein, 1999) and *equal weighting* (Dawes and Corrigan, 1974, Einhorn and Hogarth, 1975). One reason decision-making heuristics are based on the idea that good reasoning often requires just a single piece of information (Marewski et al., 2010). Applying such a strategy to the initial example, you would only need to inspect the most important attribute: the movie rating. Based on this attribute, you decide to watch Movie A and ignore all other information about both movies. Equal weighting heuristics on the other hand completely abstain from differentiating between the attributes and instead tally all of them together to decide which option to choose. In our example, Movie B has two attributes in its favor, while Movie A only has one. Hence, you would decide to watch Movie B if your decision was based on an equal weighting heuristic.

Even though they are computationally simplistic strategies, heuristics can be surprisingly competitive on real-world benchmarks (Czerlinski et al., 1999, Lichtenberg and Şimşek, 2017). This observation led different researchers to consider heuristics as *ecologically rational* strategies (Gigerenzer and Todd, 1999, Gigerenzer and Gaissmaier, 2011, Payne et al., 1993), implying that heuristics are strategies that are particularly well-suited for our complex and dynamic world. The ecological rationality of heuristics also makes it appealing to view them as models of human decision-making. Empirical studies attempting to show that people actually apply heuristics have, however, produced mixed evidence (Ayal and Hochman, 2009, Bröder, 2000, Glöckner and Betsch, 2008, Bröder and Gaissmaier, 2007, Hilbig, 2010, see also our later discussion on empirical results).

In this chapter, we apply BMI to decision-making problems like the aforementioned movie example; i.e., we set it up to infer decision-making strategies. Like ideal observer models, BMI attempts to infer optimal decision-making strategies but does so while taking computational resources into account. Like heuristics, strategies inferred through BMI are tailored to a specific environment. However, unlike heuristics, the inductive biases of such strategies have been meta-learned through previous interactions with the environment instead been built-in by design.

Through a series of model simulations, we demonstrate that BMI discovers several previously suggested heuristics. Specifically, our results reveal three important classes of environments that lead to three different strategies. First, if the model knows the correct ranking of attributes but not their weights, then it learns a strategy that makes decisions based only on the attribute with the highest ranking, a form of one reason decision-making. Secondly, if the model knows that the direction of correlation between attributes and outcome is positive, then it learns a strategy that makes decisions based on equal weighting. Finally, if the model does not know either the ranking or the direction of attributes, it learns to use individual weights for each attribute. This analysis provides new insights into the mixed results of prior empirical work on heuristics because it makes precise predictions about if and when a specific heuristic should be used. We verify these predictions in three empirical paired comparison studies and show that the vast majority of participants apply heuristics whenever they are optimal strategies for the current environment after taking limited computational resources into account.

In summary, our work makes the following three main contributions:

1. We show that heuristics can emerge through BMI, thereby providing a normative justification for previously suggested heuristics.

2. We clearly map out which features of an environment lead to which (heuristic) decision-making strategy, where knowing the correct ranking of attributes leads to one reason decision-making, knowing the directions of the attributes leads to equal weighting, and not knowing about either leads to strategies that use weighted combinations of multiple attributes.

3. We test these predictions empirically in three experiments and find strong evidence for our theory's predictions, thereby reconciling several past contradictory results.

The remainder of the chapter is organized as follows: we first summarize the relevant literature on heuristic decision-making and introduce its general terminology. Afterwards, we present formal models corresponding to different hypotheses considered in our work. By running simulations on different environments, we generate several predictions of our theory, which we empirically test in three new decision-making experiments. Finally, we discuss our results and connect our theory to related ideas.

4.2 Past Research on Heuristic Decision-Making

There has been an extensive amount of past research on heuristic decision-making. In this section, we describe common heuristics and review prior studies in the paired comparison setting with a focus on the empirical evidence they provide for heuristic decision-making.

4.2.1 Heuristics Toolbox

Although a mathematically precise definition of what constitutes a heuristic is still a topic of ongoing debates (Van Rooij et al., 2012, Chater et al., 2003), here we adopt the following definition put forward by Gigerenzer and Gaissmaier (2011): “A heuristic is a strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods.”

The collection of different heuristics is often thought of as an adaptive toolbox from which appropriate decision-making strategies can be selected as required (Gigerenzer and Selten, 2002). We are primarily interested in heuristics that can be applied to paired comparison tasks (e.g., Martignon and Hoffrage, 2002) like the aforementioned movie example. In such tasks, a decision-making agent is asked to judge which of two options is superior on an unobserved criterion. To aid the decision-making process, the agent observes multiple attributes of both options, also known as cues or features in the decision-making literature. Most heuristics developed for the paired comparison setting make use of binary features that indicate whether an attribute is present or not.*

Many heuristics are built around the concept of feature validity (Todd and Dieckmann, 2005). The validity of a binary feature is the rate at which it allows the agent to make correct predictions given that the feature is present in one option but not the other (Lee and Cummins, 2004). For example, the validity of being directed by a known director for predicting whether you like a movie

*Note that non-binary features, like average movie ratings, can always be dichotomized at a loss of information. In past studies, this has been frequently done by setting values which were less than the median to 0 and otherwise to 1.

could be 0.8, indicating that you would enjoy a movie that is directed by someone you know over someone you do not in eighty percent of the cases. In general, decision-making strategies for paired comparison tasks can be categorized into two classes: compensatory and non-compensatory strategies. A strategy is compensatory whenever it integrates information from multiple features, whereas it is non-compensatory when a feature cannot be outweighed by any combination of less important features (Rieskamp and Hoffrage, 1999).

The *weighted additive strategy* (WADD, Gigerenzer and Goldstein, 1996) is an example for compensatory decision-making. WADD weights features by their validities and decides for the option with the larger sum of weighted features. Although WADD combines information from multiple sources, it is – according to our definition – a heuristic, because it ignores potential interactions between features. In our movie example, this would correspond to weighting and adding all features together without paying attention to how they might interact (e.g., a movie database could potentially always dislike Oscar-winning movies for being too mainstream; WADD would ignore this interaction).

Most heuristics are, however, much simpler than WADD. Equal weighting heuristics, for example, are compensatory, yet simple, decision-making strategies. They do not distinguish between how features are weighted and instead use an identical weighting for all features. The process itself is realized by tallying features of both options together and selecting the one with the larger sum (Dawes and Corrigan, 1974, Einhorn and Hogarth, 1975).

The prime example for a non-compensatory strategy is the *take-the-best heuristic* (Gigerenzer and Goldstein, 1996, TTB,). TTB belongs to the family of one reason decision-making heuristics. It assumes a ranking of features based on their validities and inspects features in decreasing order until a feature that discriminates between both options is reached. The final decision is based on the validity of that feature alone, ignoring all other information. Should a ranking of features not be a priori accessible, then it can either be estimated from observations or a random ranking can be used. A TTB strategy using a random ranking of features is referred to as the Minimalist heuristic (Gigerenzer and Goldstein, 1996).

Given the seemingly endless pool of decision-making strategies, one might ask: how do people decide which strategy to apply? This problem is known as the strategy selection problem. An influential theory on how people solve the strategy selection problem is that they select the strategy that works best in a particular environment (Erev and Barron, 2005, Rieskamp and Otto, 2006). In a series of papers Hogarth and Karelaia (2006, 2005, 2007) provided guidance for the selection of strategies by characterizing environmental conditions under which different heuristics – like TTB and equal weighting – are equivalent to ideal observer models in terms of their performance. Their results suggest that there exist environments in which even a fully rational decision-maker could opt to use heuristics.

4.2.2 Empirical Studies

The observation that heuristics are computationally efficient and ecologically rational strategies is often used to justify them as models of human decision-making (Todd and Gigerenzer, 2007). However, to truly establish that people actually use heuristics, proving good performance in simulation is not sufficient; it also requires empirical evidence. Many studies have attempted to find such

Paper	Learning	Cost	Tasks	Trials/Task	Options	Features	Discretized	Ranking	Direction	Evidence
Rieskamp and Otto (2006, Study 1)	✗	🖱	1	168	2	6	✓	✓	+	✓
Rieskamp and Otto (2006, Study 2)	✓	🖱	1	182	2	6	✓	✗	+	✓
Glöckner and Betsch (2008, Study 2)	✗	🖱	1	138	3	3	✓	✓	+	✓
Scheibehenne et al. (2013)	✗	🖱	1	48	2	6	✓	✓	+	✗
Van Ravenzwaaij et al. (2014, Study 1)	✗	🖱	1	100	2	9	✓	✓	+	✗
Van Ravenzwaaij et al. (2014, Study 2)	✗	🖱	1	100	2	9	✓	✓	+	✗
Bröder (2000, Study 3)	✗	$	1	120	2	4	✓	✓	+	✓
Rieskamp and Otto (2006, Study 3)	✗	$	1	168	2	6	✓	✓	+	✓
Dieckmann and Rieskamp (2007)	✗	$	1	96	2	6	✓	✓	+	✓
Newell et al. (2003, Study 1)	✓	$	1	60	2	6	✓	✓	+	✗
Newell et al. (2003, Study 2)	✓	$	1	60	2	2	✓	✓	+	✗
Newell and Lee (2011, Study 2)	✓	$	1	80	2	6	✓	✗	+	✗
Newell et al. (2007, Study 1)	✓	💾	1	102	2	4	✓	✗	?	✓
Bröder and Schiffer (2003)	✗	💾	1	52	10	4	✓	✓	?	✓
Bröder and Schiffer (2006)	✗	💾	1	52	10	4	✓	✓	?	✓
Bröder and Gaissmaier (2007)	✗	💾	1	52	10	4	✓	✓	?	✓

Table 4.1: Empirical studies that involved costs to acquire information about features. The learning column indicates whether validities/weights were provided (✗) or had to be learned (✓). The cost column describes how costs to acquire features were implemented, with 🖱 referring to a Mouselab paradigm, $ indicating that monetary fees are required to reveal a feature and the floppy disk representing a memory-based retrieval setting. The direction column shows the direction of features, with + for positive directions and ? for unknown directions. The evidence column indicates whether the study found evidence for heuristics (✓) or not (✗).

evidence, yet no consensus for or against heuristics has been reached. Here, we provide an overview of these studies and attempt to connect their findings. We consider studies in which information about features was freely accessible and those that included a cost for obtaining information. Tables 4.1 and 4.2 summarize characteristics of studies with and without costs to acquire information, respectively.

Problems where it is costly to access feature values naturally favor strategies

Paper	Learning	Cost	Tasks	Trials/Task	Options	Features	Discretized	Ranking	Direction	Evidence
Bergert and Nosofsky (2007, Study 1)	✓	✗	1	160	2	6	✓	✗	+	✓
Bergert and Nosofsky (2007, Study 2)	✓	✗	1	160	2	6	✓	✗	+	✓
Bröder (2000, Study 1)	✓	✗	1	30	2	5	✓	✗	+	✗
Bröder (2000, Study 2)	✓	✗	1	120	2	5	✓	✗	+	✗
Lee and Cummins (2004)	✓	✗	1	5	2	6	✓	✗	+	✗
Glöckner and Betsch (2008, Study 1)	✗	✗	1	138	3	3	✓	✓	+	✗
Newell and Lee (2011, Study 1)	✓	✗	1	40	2	6	✓	✗	+	✗
Parpart et al. (2017)	✓	✗	1	10	2	4	✓	✗	+	✗
Newell et al. (2009, Study 1)	✓	✗	1	240	2	4	✓	✗	?	✗
Newell et al. (2009, Study 2)	✓	✗	1	240	2	4	✓	✗	+	✓
Gluck et al. (2002)	✓	✗	1	200	2	4	✓	✗	?	✓
Lagnado et al. (2006)	✓	✗	1	200	2	4	✓	✗	?	✗
Juslin et al. (2003a)	✓	✗	1	130	2	4	✓	✗	?	✗
This work (Study 1)	✓	✗	30	10	2	4	✗	✓	?	✓
This work (Study 2)	✓	✗	30	10	2	4	✗	✗	+	✓
This work (Study 3)	✓	✗	30	10	2	4	✗	✗	?	✗

Table 4.2: Empirical studies that involved no costs to acquire information about features. The learning column indicates whether validities/weights were provided (✗) or had to be learned (✓). The direction column shows the direction of features, with + for positive directions and ? for unknown directions. The evidence column indicates whether the study found evidence for heuristics (✓) or not (✗).

that only require a few pieces of information. Because of that, studies in this context concentrated on one reason decision-making heuristics such as TTB. For our review, we look at studies in the Mouselab paradigm (Payne et al., 1988), studies with monetary costs (Newell et al., 2003) and memory-based retrieval studies (Bröder and Schiffer, 2003).

The Mouselab paradigm is a process-tracing approach to decision-making, which requires participants to click or hover over a specific feature to reveal its value. This paradigm allows researchers to identify which information is considered by the participant. In studies making use of the paradigm, Rieskamp and Otto (2006) showed that people's selection of strategies depended on the environment they interacted with. Participants in their study had initial preferences for compensatory strategies, but then slowly adopted TTB in a non-compensatory environment and WADD in a compensatory one. However, other studies with comparable conditions arrived at different conclusions. For example, Scheibehenne et al. (2013) demonstrated that people were better described through a mixture of TTB and WADD even in non-compensatory environments, indicating a general preference for compensatory strategies. Van Ravenzwaaij et al. (2014) showed that hierarchical models accounting for both search order and termination provided a better explanation for participants' choices than TTB and WADD.

Requiring a monetary cost to reveal features is another process-tracing approach. Like the Mouselab paradigm, it facilitates strategies that rely on less information. In several experiments with monetary costs, Bröder (2000) produced evidence in favor of one reason decision-making heuristics. In his experiments, more participants were classified as TTB users in a high-cost condition compared to a low-cost condition. Similarly, Dieckmann and Rieskamp (2007) observed that TTB predicted more decisions in environments with monetary costs. However, Newell et al. (2003) demonstrated that even with large monetary costs and other conditions favoring one reason decision-making heuristics, not many participants acted according to TTB.

Requiring participants to recall features from memory is yet another method to constrain the amount of information they use. In multiple experiments with memory-based retrieval, Bröder and colleagues demonstrated that participants became more consistent with TTB when features had to be retrieved from memory (Bröder and Schiffer, 2003, 2006, Bröder and Gaissmaier, 2007). Bröder and Schiffer (2003), for example, classified 72% of participants as TTB users when they were under high working memory load, but only 56% when they were not.

In general, studies with increased costs for utilizing information indicate that human decision-making becomes more consistent with one reason decision-making heuristics. Nonetheless, even under supposedly favorable conditions, prior research did not reach a clear consensus on whether people use one reason decision-making heuristics or if they rely on more complex strategies instead.

Glöckner and Betsch (2008) argued that process-tracing studies are likely to underestimate the cognitive capacity of participants, as they hinder the activation of automatic decision-making processes. They verified this claim by demonstrating that participants were generally able to combine information from multiple features extremely quickly when the acquisition of information was not constrained. Further studies with freely accessible information provided similar results (Bröder, 2000, Lee and Cummins, 2004, Parpart et al., 2018), always concluding that few participants made decisions consistent with TTB and that their choices were, in general, better described through compensatory strategies such as logistic regression. Newell and Lee (2011) highlighted large inter-individual differences and presented a sequential sampling model providing better fits than TTB, WADD, and a strategy selection model across all participants. Bergert and Nosofsky (2007) were among the few who provided support for heuristics in human decision-making even when information is free. They showed that people exhibit non-compensatory patterns of decision-making, assigning over half of the total weight to a single feature, and provided additional evidence for frugal strategies in form of reaction times.

To summarize, many past paired comparison studies attempted to produce evidence for one reason decision-making, thereby focusing less on other heuristics such as equal weighting. Many of them concluded that heuristic strategies were indeed more apparent when it was costly to access information. Evidence for heuristics in human decision-making in the unimpeded setting is, however, rare. Looking at Tables 4.1 and 4.2, we observe that the majority of prior empirical studies evaluated their hypothesis in environments that either explicitly or implicitly assumed positive directions of features. While it is always possible to code features such that they have positive directions (e.g., changing the feature "won an Oscar" to "did not win an Oscar" if winning an Oscar has a negative correlation with the outcome), doing so can influence the strategies people apply. To foreshadow our results, we demonstrate that a restriction to environments with known positive attribute directions causes equal weighting heuristics to become optimal under limited computational resources. Therefore, at least some of the mixed results of prior studies can be explained by the use of environments that favor strategies not considered in their analyses.

There are a number of research areas that use experimental paradigms similar to paired comparison studies and that have, interestingly enough, also produced mixed evidence on whether people rely on heuristic decision-making or not. In *probabilistic category learning* (Ashby and Maddox, 2005) participants are asked to classify objects into one of usually two categories. Thus, similar to paired comparison tasks participants learn a mapping between features and a binary outcome. Juslin et al. (2003b) noted that category learning emphasizes exemplar models, which is in contrast to the linear additive cue-integration models studied in the decision-making domain. Based on this observation they investigated which factors modulate a shift from exemplar models to cue-integration models. However, they did not examine the role of heuristic decision-making strategies in the context of category learning. In a follow-up study Juslin et al. (2003a) did consider the possibility for one reason decision-making heuristics but found little evidence for such strategies. Adding additional time pressure did not change their conclusion that most participants integrated information from multiple features, either through exemplar or cue-integration models.

Another closely connected paradigm with a long history on its own is *multiple-cue probability learning* (MCPL, Hammond, 1955, Brehmer, 1979, Gluck and Bower, 1988). In MCPL people have to learn about an imperfect relationship between an object described by multiple features and an outcome. A popular instance of MCPL is given by the weather prediction task. Here, participants are presented with a multi-dimensional stimulus taking the form of tarot cards and learn based on feedback whether given patterns lead to sunny or rainy weather. Gluck et al. (2002) conjectured that people approach this task using three different strategies: (1) an optimal strategy, in which people learn about all available features, (2) a one reason decision-making heuristic, in which decisions are based on a single feature, and (3) a singleton heuristic, in which people learn only about the patterns that have a single feature present. In two studies they found that a majority of participants (85% across both studies) was overall best fit by the singleton heuristic. As more data was observed participants either switched towards the one reason decision-making heuristic in a more challenging experiment or the optimal multi-cue strategy in an easier experiment. In a similar setup but using a different analysis, Lagnado et al. (2006) instead concluded that a vast majority of participants was best described by a strategy that integrated information from all features (86% across three studies). Newell et al. (2007) reported similar results, with the additional observation that people switched towards a more simplistic singleton heuristic if they were put under working memory load. Finally, it is worth pointing out that equal weighting heuristics also received some attention in the MCPL literature: when participants were provided with directional information about features, they switched from a multi-cue strategy towards an equal weighting heuristic (Newell et al., 2009). In the context of this article, this is an interesting observation, because – as we will show later on – it is exactly what our meta-learning models predict.

At the heart of ecologically rational heuristics is a powerful idea: there are much simpler ways of interacting effectively with many natural problems than the use of complex strategies. Moreover, computational constraints render it necessary that people should make use of these shortcuts extensively. In our summary of previous empirical studies, we have seen that this is indeed sometimes observed but certainly not always. In general, prior empirical work has produced mixed evidence on whether people do in fact rely on heuristic decision-making strategies or not. An important aspect that has been largely neglected in prior empirical work is the identification of conditions under which a particular heuristic occurs.

4.3 Computational Models

When applied to decision-making problems, BMI assumes that people make environment-specific inferences about which strategies to use, while also making optimal use of limited computational resources. Having access to such a model does allow us to predict if and when people should rely on heuristic decision-making strategies, assuming that they use available computational resources efficiently. To test this conjecture, we also introduce several other computational models of decision-making in paired comparison tasks. First, we will outline the assumptions about the structure of the problem to be solved and define a corresponding ideal observer model. Then, we will introduce probabilistic variants of two popular heuristics. Both heuristics are considerable simplifications with respect to how they use information compared to the ideal observer model. Finally, we will describe how BMI can be applied to decision-making problems.

The decision-making problems we focus on in this chapter are paired comparison tasks with continuous features. In a paired comparison task an agent – either human or machine – has to decide which of two options with feature vectors $\mathbf{x}_{A,B} \in \mathbb{R}^d$ has the higher value on an unobserved criterion $y_{A,B}$. In our movie example, the feature vector contains information about whether the movie has won an Oscar, its average rating on a reviewing website and so on, while the unobserved criterion corresponds to your personal rating of the movie (i.e., how much you would like the movie). We consider the setting where data arrives sequentially, i.e. one at a time, and with feedback that indicates which option had the higher criterion value. Let $\mathbf{x}_{A,t}$ and $\mathbf{x}_{B,t}$ denote the observed features at time-step t and let c_t be a binary variable that takes the value of 1 if option A has the higher criterion value and 0 otherwise. In each time-step, the agent first observes both options, then makes a prediction about c_t, and subsequently receives feedback about which option actually had the higher criterion value. Note that learning in this setting is always based on feedback in form of c_t, and that the unobserved criteria $y_{A,t}$ and $y_{B,t}$ are never observed directly.

In contrast to most prior work, we investigate paired comparison tasks with continuous features. In many real-world scenarios, features are naturally described through continuous values and thus we believe that the restriction to binary features neglects a characteristic present in many of the problems people typically solve. Moving to continuous features also facilitates statistical analysis as fewer trials are needed to observe expected effects. For example, it would require over four times more trials to distinguish an ideal observer model from a single cue heuristic in environments with dichotomized features instead of continuous ones (see Appendix A for further details).

4.3.1 Ideal Observer

Ideal observer (IO) models are designed to provide a theoretical upper bound on performance in a specific task. In the following, we construct an ideal observer model for paired comparison tasks. For this, we assume that there exists an underlying linear relationship between features and the criterion:

$$\begin{aligned} y_A &= \mathbf{w}^T \mathbf{x}_A + \epsilon_A \\ y_B &= \mathbf{w}^T \mathbf{x}_B + \epsilon_B \end{aligned} \tag{4.1}$$

with feature weights $\mathbf{w} \in \mathbb{R}^d$ and independent, additive noise $\epsilon_{A,B} \sim \mathcal{N}(0, \sigma^2)$. Under this assumption we can express the probability, that option A has a higher criterion value than option B as:

$$\begin{aligned} p(Y_{A,t} > Y_{B,t} | \mathbf{x}_{A,t}, \mathbf{x}_{B,t}, \mathbf{w}, m = \mathrm{IO}) &= p(C_t = 1 | \mathbf{x}_t, \mathbf{w}, m = \mathrm{IO}) \\ &= \mathbf{\Phi}\left(\frac{\mathbf{w}^T \mathbf{x}_t}{\sqrt{2}\sigma} \right) \end{aligned} \tag{4.2}$$

where $\mathbf{\Phi}$ is the cumulative distribution function of a standard normal distribution. For ease of notation, we have denoted the difference between feature vectors as $\mathbf{x}_t = \mathbf{x}_{A,t} - \mathbf{x}_{B,t}$ and used a binary random variable C_t to indicate which of the two options has a higher criterion value.

Equation 4.2 is known in the statistics and machine learning literature as *probit regression model*. The probit regression model makes it clear that an ideal observer should represent the probability that one option is better than the other using a weighted sum of differences between features of the options. Hence, the ideal observer model is a compensatory decision-making strategy.

Parameter Estimation

Equation 4.2 provides an ideal observer model under the assumption that underlying weights $\mathbf{w}$ are known. However, we assume that weights are not provided in advance to the decision-making agent. Thus, the agent has to infer them based on past observations. An ideal observer can apply Bayesian inference to infer unobserved parameters from data in a normative manner. In our setting, we estimate unobserved parameters by applying Bayesian inference sequentially.

Exact inference is not possible under the above assumptions and thus we resort to a *variational approximation* (Jordan et al., 1999). We approximate the true posterior is with a normal distribution $q(\mathbf{w}|\boldsymbol{\lambda}_t) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and optimize its parameters $\boldsymbol{\lambda}_t = (\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ through gradient ascent on the evidence lower bound:

$$\mathcal{L}(\boldsymbol{\lambda}_t) = \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\lambda}_t)} \left[\log p(C_t = c_t|\mathbf{x}_t, \mathbf{w})\right] - \mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\lambda}_t)||q(\mathbf{w}|\boldsymbol{\lambda}_{t-1})\right] \tag{4.3}$$

where $q(\mathbf{w}|\boldsymbol{\lambda}_0)$ corresponds to an initial prior distribution. This kind of approximation is equivalent to exact inference when the true posterior is within the considered variational family. We provide further details on how Equation 4.3 is optimized in Appendix B.

To make predictions, we average over all plausible parameter values given by the variational distribution. The resulting predictive distribution can be expressed in closed form:

$$p(C_{t+1} = 1|\mathbf{x}_{t+1}, \boldsymbol{\lambda}_t) = \int p(C_{t+1} = 1|\mathbf{x}_{t+1}, \mathbf{w}) q(\mathbf{w}|\boldsymbol{\lambda}_t) d\mathbf{w} = \boldsymbol{\Phi}\left(\frac{\boldsymbol{\mu}_t^T \mathbf{x}_{t+1}}{\sqrt{2\sigma^2 + \mathbf{x}_{t+1}^T \boldsymbol{\Sigma}_t \mathbf{x}_{t+1}}}\right) \tag{4.4}$$

Furthermore, we assume that features weights are sampled from a standard normal distribution at the beginning of each task and held constant over its entire duration, which implies that an ideal observer should use a prior in form of a standard normal distribution, i.e. $q(\mathbf{w}|\boldsymbol{\lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I})$.

4.3.2 Heuristics

The two heuristics we consider in our analysis belong to the categories of one reason decision-making and equal weighting. In contrast to traditional heuristics, like TTB, they are probabilistic decision-making strategies for tasks with continuous features. Both are obtained through modification of the ideal observer model, such that either less information is required to make a decision or that information is combined in a simpler way.

One Reason Decision-Making

In our implementation of one reason decision-making, we modify Equation 4.2 and replace it with a model that only takes a single feature $\mathbf{x}_t^*$ into account:

$$p(C_t = 1|\mathbf{x}_t, w, m = \text{SC}) = \mathbf{\Phi}\left(\frac{w \cdot \mathbf{x}_t^*}{\sqrt{2}\sigma}\right) \tag{4.5}$$

We refer to the resulting strategy as *single cue* (SC) heuristic. If a ranking of features is available, decisions are based on the most predictive feature, otherwise we select the feature that performed best on the data so far. In contrast to TTB, the single cue heuristic does not involve sequential search over features. However, we assume that features take continuous values, and hence search is not required as a feature nearly always discriminates between options (Luan et al., 2014).

Equal Weighting

In our probabilistic version of *equal weighting*, we replace Equation 4.2 with a model that has a single, tied weight for all features:

$$p(C_t = 1|\mathbf{x}_t, w, m = \mathrm{EW}) = \boldsymbol{\Phi}\left(\frac{w \cdot \sum_{i=1}^{d} \mathbf{x}_{t,i}}{\sqrt{2}\sigma}\right) \tag{4.6}$$

If $w > 0$, this equal weighting heuristic probabilistically selects the option with the larger sum of features. For $w < 0$, it becomes more likely to select the option with the smaller sum. Using a negative weight is appropriate if most features have negative correlations with the criterion. Note that the ideal observer model contains as many free parameters as there are observed features, while both heuristics have only a single free parameter regardless of how many features are observed.

4.3.3 Strategy Selection Model

Theories of strategy selection argue that based on repeated interactions with an environment people learn to select the strategy from a given repertoire that works best in that environment (Erev and Barron, 2005, Rieskamp and Otto, 2006). We also consider the possibility that human choices are based on a strategy selection model in our later analysis. The strategy selection model used here is based on the idea of *Bayesian model selection* (Bishop, 2006). In time-step $t + 1$, the agent selects the model m with the highest posterior probability given the past data $p(m|\mathbf{x}_{1:t}, c_{1:t})$ from a set of candidate models $\mathcal{M}$. In our model simulations we define the set of candidate models as $\mathcal{M} = \{\mathrm{IO}, \mathrm{SC}, \mathrm{EW}\}$ and assume a uniform prior over models. The computation of the posterior distribution over models can be expressed recursively:

$$\begin{aligned}
m_{t+1} &= \arg\max_{m \in \mathcal{M}} \left[\log p(m|\mathbf{x}_{1:t}, c_{1:t})\right] && (4.7)\\
&= \arg\max_{m \in \mathcal{M}} \left[\log p(c_{1:t}|\mathbf{x}_{1:t}, m) + \log p(m)\right] && (4.8)\\
&= \arg\max_{m \in \mathcal{M}} \left[\log p(c_{1:t}|\mathbf{x}_{1:t}, m)\right] && (4.9)\\
&= \arg\max_{m \in \mathcal{M}} \left[\log p(c_t|\mathbf{x}_t, \mathbf{x}_{1:t-1}, c_{1:t-1}, m) + \log p(c_{1:t-1}|\mathbf{x}_{1:t-1}, m)\right] && (4.10)\\
&= \arg\max_{m \in \mathcal{M}} \left[\log p(c_t|\mathbf{x}_t, \boldsymbol{\lambda}_{t-1}, m) + \log p(c_{1:t-1}|\mathbf{x}_{1:t-1}, m)\right] && (4.11)
\end{aligned}$$

Equation 4.11 reveals that this strategy selection model amounts to selecting the model with the highest accumulated log-evidence over all previous time-steps. The strategy selection model combines advantages of the ideal observer model with those of heuristics: if additional information is provided heuristics may outperform the ideal observer early on and hence they will be initially preferred. After a while, the ideal observer model surpasses both heuristics in terms of performance and hence it will be preferred during later stages of a task.

4.3.4 Bounded Meta-Learned Inference

Next, we explain how one can meta-learn an algorithm that infers decision-making strategies in a paired comparison task. The idea is simple: instead of using Bayesian (or variational) inference to infer posterior distributions over probit regression weights, we train a RNN to make this inference. In time-step $t + 1$ the network processes the previous feature vector $\mathbf{x}_t$ together with its corresponding target c_t and uses this information to update its hidden state $\mathbf{h}_t$. The parameters of the posterior distribution $\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\}$ are then computed through a linear transformation of the hidden state:

$$\boldsymbol{\mu}_t = \mathbf{W}_{\boldsymbol{\mu}}\mathbf{h}_t \tag{4.12}$$

$$\log \boldsymbol{\sigma}_t = \mathbf{W}_{\boldsymbol{\sigma}}\mathbf{h}_t \tag{4.13}$$

$$\boldsymbol{\Sigma}_t = \text{diag}\left(e^{\log \boldsymbol{\sigma}_t}\right) \tag{4.14}$$

Finally, the model combines the estimated weights with the feature vector $\mathbf{x}_{t+1}$ as described in Equation 4.4 to obtain the predictive posterior distribution:

$$p(C_{t+1} = 1|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, c_{1:t}, \boldsymbol{\Theta}) = p(C_{t+1} = 1|\mathbf{x}_{t+1}, \boldsymbol{\lambda}_t, \boldsymbol{\Theta}) \tag{4.15}$$

$$= \int p(C_{t+1} = 1|\mathbf{x}_{t+1}, \mathbf{w})q(\mathbf{w}|\boldsymbol{\lambda}_t)d\mathbf{w} \tag{4.16}$$

Figure 4.1 illustrates graphically how the RNN processes a sequence of observations. The described setup deviates slightly from the one presented in Section 3.1. Instead of estimating the predictive posterior distribution directly, we have chosen to make the dependence probit regression weights explicit. Doing so makes it easier to analyze what kind of strategies BMI infers. The general concepts, however, remain the same.

Initially, the RNN implements a random mapping. During meta-learning it is then turned into a resource-rational learning algorithm. This is accomplished by

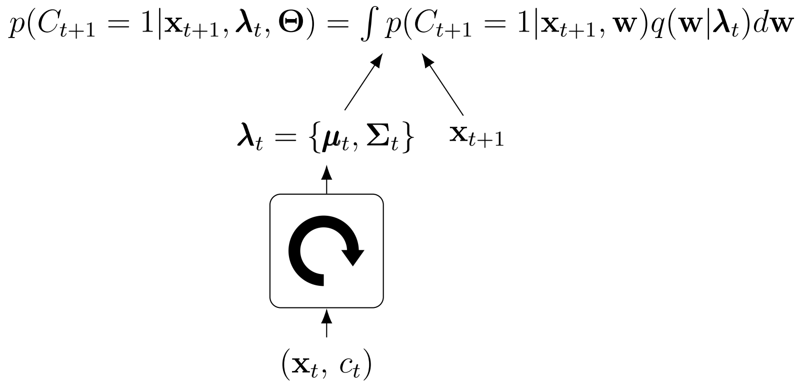

$p(C_{t+1} = 1|\mathbf{x}_{t+1}, \boldsymbol{\lambda}_t, \boldsymbol{\Theta}) = \int p(C_{t+1} = 1|\mathbf{x}_{t+1}, \mathbf{w}) q(\mathbf{w}|\boldsymbol{\lambda}_t) d\mathbf{w}$
$\boldsymbol{\lambda}_t = \{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\}$
$\mathbf{x}_{t+1}$
$(\mathbf{x}_t,\, c_t)$

Figure 4.1: Graphical depiction of BMI for paired comparison tasks. The RNN sequentially processes examples from a given task. Through its recurrent activations it combines information from all previous feature-target pairs to compute a distribution over weights, which is then combined with the next input to obtain the predictive distribution.

minimizing the BMI objective until convergence:

$$\mathcal{L}_{\text{BMI}}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\mathbb{E}_{p(\mathbf{x}_{1:T}, c_{1:T})}\left[\sum_{t=0}^{T-1} - \log p(C_{t+1} = c_{t+1}|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, c_{1:t}, \boldsymbol{\Theta})\right]\right] \\ + \beta \text{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})||p(\boldsymbol{\Theta})\right] \tag{4.17}$$

After meta-learning is completed, the RNN acts as an environment-specific learning algorithm that makes optimal use of limited computational resources. In Appendix C we provide a full specification on the network architecture, meta-learning procedure and choice of prior.

At this point, it seems sensible to ask: what types of decision-making strategies can BMI infer? Both the single cue heuristic and equal weighting are subsets of the space of all possible weight vectors that can be inferred. Equal weighting heuristics correspond to uniform vectors (e.g., $[1, 1, 1, 1]$), while single cue heuristics can be expressed through a vector with a single non-zero entry (e.g., $[1, 0, 0, 0]$). BMI could thus – in principle – discover the two heuristics and select between them whenever appropriate. We know that MI – or equivalently BMI with $\beta = 0$ – approximately simulates an ideal observer, and hence we expect it to infer strategies that use independent and non-zero weights for all features. However, as we decrease the description length of the emerging learning algorithm, we expect it to infer simpler strategies like the single cue or equal weighting heuristic. Importantly, which strategy BMI infers, and whether it corresponds to a particular heuristic or not, does not only depend on its complexity but also on the distribution over tasks that was used for meta-learning.

4.3.5 Feedforward Network

We also compare our models to a simple *feedforward neural network* baseline model. The feedforward network uses the same architecture as MI and BMI, but without recurrent connections and without the previous target as additional input. Learning is performed through gradient descent on the negative log-probabilities of observed targets. The exact forward pass equations are given by:

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_{ir}\mathbf{x}_t\right) \tag{4.18}$$

$$\mathbf{z}_t = \sigma\left(\mathbf{W}_{iz}\mathbf{x}_t\right) \tag{4.19}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \tanh\left(\mathbf{W}_{ih}\mathbf{x}_t + \mathbf{r}_t\right) \tag{4.20}$$

$$\boldsymbol{\mu}_t = \mathbf{W}_{\boldsymbol{\mu}}\mathbf{h}_t \tag{4.21}$$

$$\log \boldsymbol{\sigma}_t = \mathbf{W}_{\boldsymbol{\sigma}}\mathbf{h}_t \tag{4.22}$$

$$\boldsymbol{\Sigma}_t = \operatorname{diag}\left(e^{\log \boldsymbol{\sigma}_t}\right) \tag{4.23}$$

where σ denotes the logistic sigmoid function and $\odot$ element-wise multiplication.

4.3.6 Model Summary

Let us briefly summarize all outlined models again and contrast the assumptions they make:

Ideal observer model Assumes that everything about the structure of the decision-making environment is known. Specifically, it knows about the linear-Gaussian relationship. With this knowledge, it is able to compute the optimal solution by combining information from all features through weighted sums.

Heuristics Assume that computing weighted sums is too burdensome and instead bet on simpler ways for making decisions, like using only a single feature or using an equal weight for all features.

Strategy selection model Assumes a predefined repertoire of strategies and selects the one that works best on a given task after repeated interactions with it.

BMI Does not know anything about the structure of the environment explicitly. Instead, it uses a resource-rational algorithm that has been acquired through repeated encounters with the environment to infer decision-making strategies.

Feedforward network Based on same non-linear model architecture as BMI. However, learning is implemented through gradient descent instead of the forward dynamics of a RNN. The learning algorithm is not adapted to the environment and there are no resource limitations.

4.4 Model Simulations

Next, we demonstrate through a series of model simulations that BMI recovers both single cue and equal weighting heuristics in specific environments. This implies that both heuristics can be resource-rational strategies under certain conditions. However, we also identify circumstances where BMI does not discover any known heuristic and instead infers strategies that use weighted combinations of all features. Before running these simulations, we first have to specify the assumptions we make about the environment and introduce a method for analyzing the emerging strategies.

4.4.1 Environments

Applying BMI to decision-making problems requires to specify a distribution over tasks $p(\mathbf{x}_{1:T}, c_{1:T})$ that is used for meta-learning. In general, this distribution should reflect a participant's prior experiences in the world and its expectations about what tasks might be encountered during the experiment. Here, we make the following assumptions. To generate a single task, we proceed in three steps:

1. Randomly generate features $\mathbf{x}_{A,t}$ and $\mathbf{x}_{B,t}$ from a multivariate normal dis-

tribution with zero mean and a given covariance matrix. All tasks presented in this chapter involve four-dimensional feature vectors.

2. Randomly generate features weights (ref. Equation 4.1 or 4.2) by sampling from a standard normal distribution.

3. Randomly determine which option has the larger criterion by sampling from a Bernoulli distribution with a success probability given by Equation 4.2.

Features weights are held constant over a task but are resampled between tasks. Importantly, we assume that the decision-making agent cannot access these weight vectors directly, but instead has to infer them based on observations. An unrestricted meta-learned algorithm that is trained on such an environment will be approximately equivalent to our ideal observer model.

Both redundancy and uncertainty are crucial factors in many real-world decision-making problems (Gigerenzer and Gaissmaier, 2011). Thus, we want them to be present in our environments. Partially redundant features are ensured by drawing separate feature covariance matrices from a LKJ prior with $\eta = 2$ (Lewandowski et al., 2009) for each task. To introduce uncertainty, we use a limited number of trials in each task ($T = 10$) and set the additive noise term σ such that an ideal observer is correct in 85% of the cases in the tenth trial.

We consider three variations of the previously outlined environments, that assume (1) known rankings of features, (2) known directions of features, or (3) neither. To provide agents with a ranking of features, we arrange them in decreasing order according to the magnitude of their weights. Known directions are ensured by inverting the sign of a feature if it has a negative correlation with the criterion, leading to features with only positive directions. Note that our ideal observer implementation always assumes the original standard normal prior over weights, i.e. the prior is not adjusted based on the additional information about ranking or direction. These environments are used during meta-learning, for the model simulation results presented next, and to generate the tasks for our empirical studies.

4.4.2 Strategy Analysis

To characterize different decision-making strategies, we adopt a measure from the economics literature called the *Gini coefficient* (Atkinson et al., 1970). The Gini coefficient was originally intended to describe income and wealth distributions of countries. Its minimal value of zero corresponds to a country in which all residents are equally wealthy, while the maximal value of one corresponds to a country in which a single person possesses everything.[†]

The extreme cases of the Gini coefficient also coincide with the two previously discussed heuristics: equal weighting heuristics have a Gini coefficient of zero, while single cue heuristics have a Gini coefficient close to one. Thus, we can employ the Gini coefficient to understand how similar estimated regression weights are compared to both heuristics. In practice, we compute Gini coefficients from absolute values of weight vectors. Mathematically, the Gini coefficient of a weight vector $\mathbf{w} \in \mathbb{R}^d$ is defined as half of the relative mean absolute

[†] The extreme value of one is only reached in the limit of an infinite number of residents, otherwise the maximum Gini coefficient for d residents is $1 - d^{-1}$.

difference:

$$G(\mathbf{w}) = \frac{\displaystyle\sum_{i=1}^{d}\sum_{j=1}^{d} |\mathbf{w}_i - \mathbf{w}_j|}{\displaystyle 2d\sum_{i=1}^{d} \mathbf{w}_i} \tag{4.24}$$

Throughout this section, we analyze Gini coefficients for BMI (with $\beta = 0.01$), MI, and ideal observer models. If Gini coefficients are consistently close to zero or one, we deduce that the model has recovered one of the two heuristics.

Additionally, we evaluate the average KL divergence from the predictive posterior distribution of both heuristics to the predictive posterior distribution of BMI. This KL divergence can be interpreted as a distance measure between two models. If it is significantly lower for one of the two heuristics, this would further strengthen our claim that BMI has discovered that particular heuristic.

4.4.3 Model Simulations

First, we considered an environment with known feature rankings. For MI and BMI we optimized meta-parameters until convergence in an environment where features are ordered based on the magnitude for their associated weight. We then analyzed the Gini coefficients of inferred regression weights after meta-learning is completed. Because MI and BMI are adapted to the environment, they could exploit the additional ranking information to adjust how they infer strategies.

Figure 4.2 (a) visualizes Gini coefficients obtained from BMI. We observe strategies with nearly maximum Gini coefficients, which correspond to weight

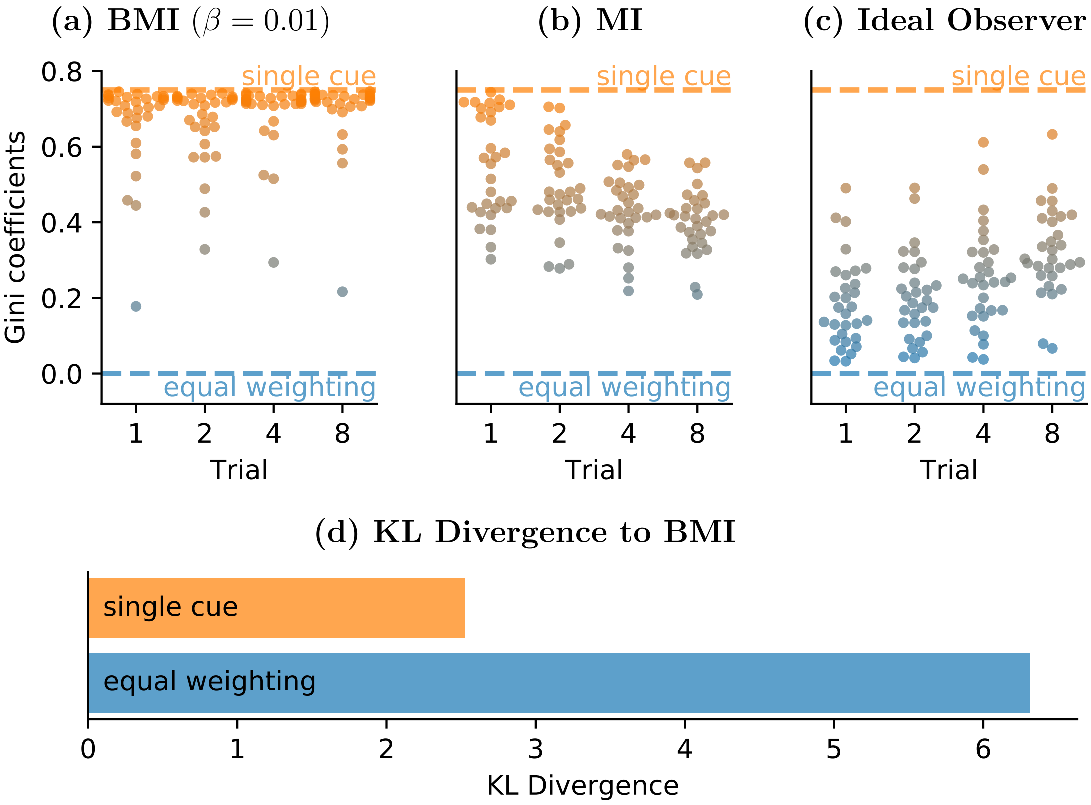


Figure 4.2: (a) to (c) Gini coefficients for an environment with known rankings. High values indicate similarity to the single cue heuristic, while low values correspond to equal weighting heuristics. (a) BMI results in Gini coefficients that are close to the single cue heuristic. (b) MI shows tendencies towards the single cue heuristic, especially with few observations. (c) Gini coefficients of the ideal observer model cover the whole range of possible values, indicating that a weighted combination of multiple features is used. (d) Average KL divergence from the predictive posterior distribution of both heuristics to the predictive posterior distribution of BMI. The KL divergence is lower for the single cue heuristic, which confirms our results from the Gini coefficient analysis.

vectors that only have a single non-zero component. Thus, we conclude that the single cue heuristic emerged as the resource-rational strategy for an environment with known feature rankings. Looking at MI in Figure 4.2 (b), we find Gini coefficients that cover a much wider range of values. Even though there is an initial tendency towards single cue heuristics, many later decisions are based on compensatory rules. This indicates that being adapted to the environment alone is not a sufficient justification for heuristics. Instead, we need algorithms that are adapted to the environment *and* efficient in terms of their computational resources. Decisions in the ideal observer model are nearly always based on weighted combinations of multiple features, and hence its Gini coefficients in Figure 4.2 (c) spread over an even wider range of values. Figure 4.2 (d) confirms our findings by showing that BMI infers predictive posterior distributions that are much more similar to the single cue heuristic than to equal weighting in terms of their KL divergence.

Next, we looked at environments where feature directions are known instead of their ranking. For this, we optimized MI and BMI in an environment with only positive feature directions. The result here looks very different compared to the ranking condition. Gini coefficients resulting from BMI, visualized in Figure 4.3 (a), are consistently close to zero. Low Gini coefficients correspond to uniform weight vectors and hence in this environment the equal weighting heuristic turned out to be rational under limited computational resources. Figure 4.3 (b) confirms earlier results showing that MI only leads towards an initial tendency towards heuristics. Early strategies are similar to equal weighting, but especially as more data is observed strategies with higher Gini coefficients emerge. The ideal observer model on the other hand does not exploit environmental charac-

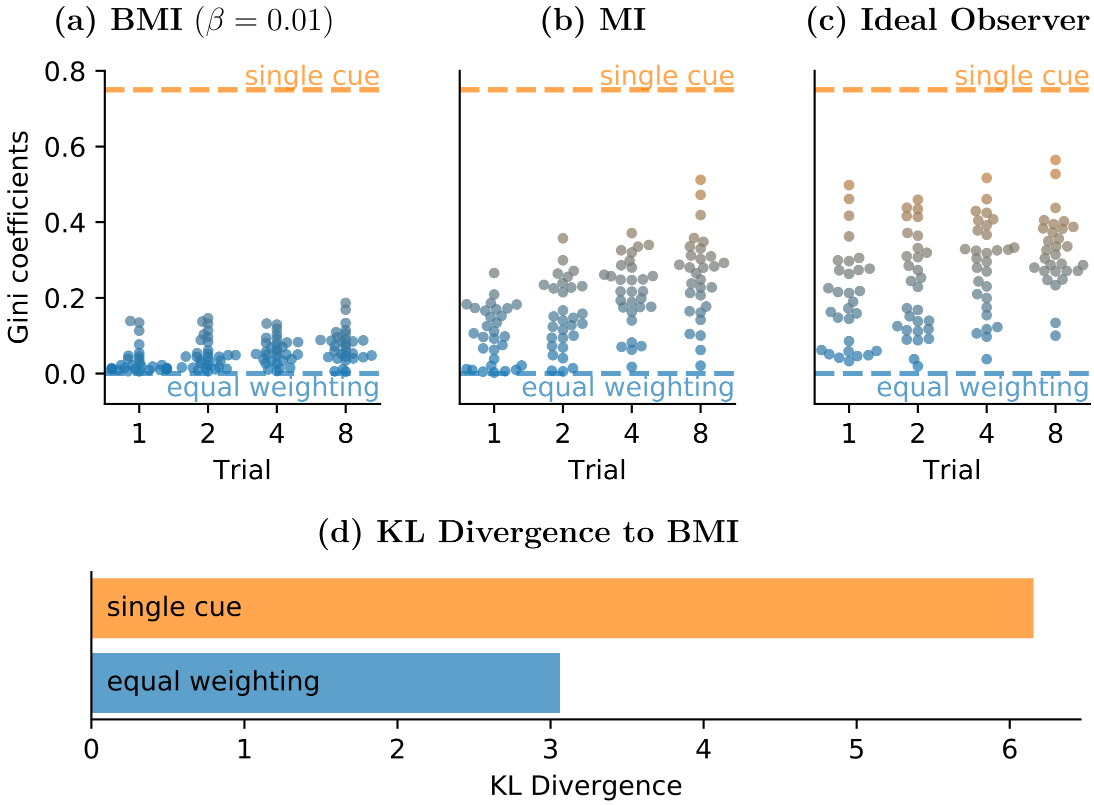


Figure 4.3: (a) to (c) Gini coefficients for an environment with positive directions. High values indicate similarity to the single cue heuristic, while low values correspond to equal weighting heuristics. (a) BMI results in Gini coefficients that are close to the equal weighting. (b) MI shows tendencies towards the equal weighting heuristic, especially with few observations. (c) Gini coefficients of the ideal observer model cover the whole range of possible values, indicating that a weighted combination of multiple features is used. (d) Average KL divergence from the predictive posterior distribution of both heuristics to the predictive posterior distribution of BMI. The KL divergence is lower for the equal weighting heuristic, which confirms our results from the Gini coefficient analysis.

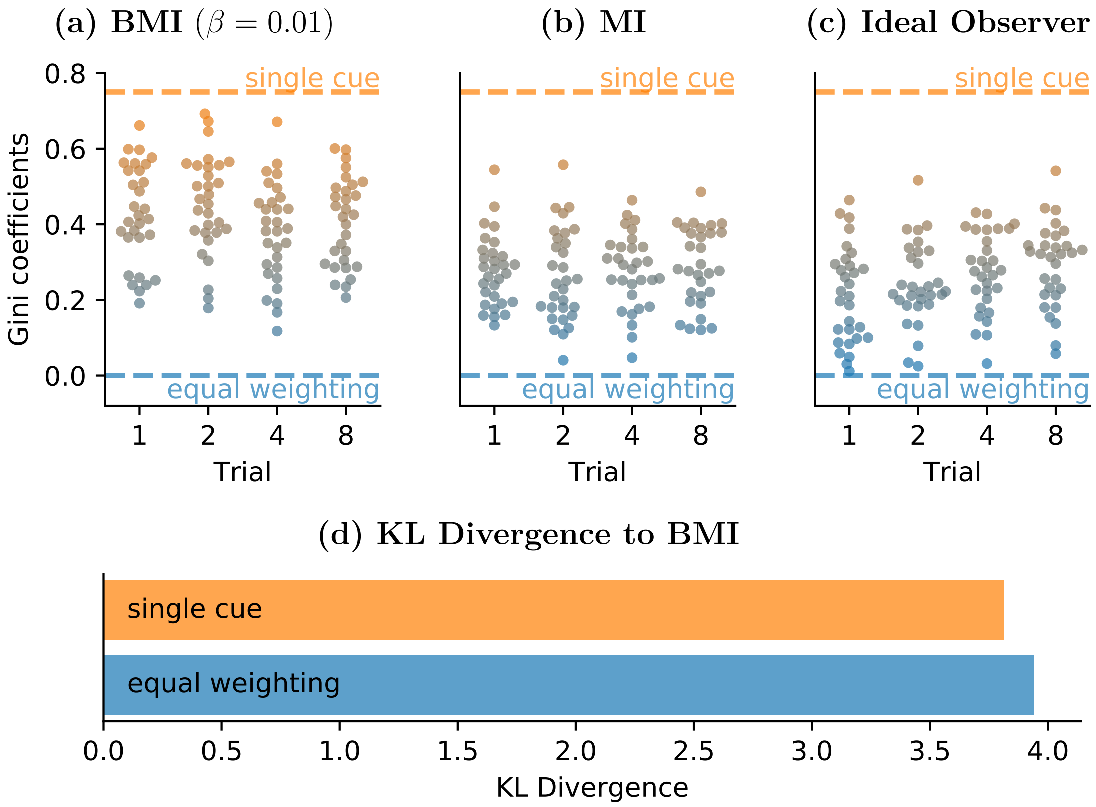


Figure 4.4: (a) to (c) Gini coefficients for an environment without ranking or direction. High values indicate similarity to the single cue heuristic, while low values correspond to equal weighting heuristics. (a) BMI, (b) MI and (c) ideal observer models result in Gini coefficients that cover the whole range of possible values, indicating that a weighted combination of multiple features is used. (d) Average KL divergence from the predictive posterior distribution of both heuristics to the predictive posterior distribution of BMI. The KL divergence is roughly equal for both heuristics, indicating that neither of the two is particularly similar to BMI.

teristics and hence we find no noticeable change in Gini coefficients compared to an environment with known rankings (Figure 4.3 (c)). As before, our results are confirmed when looking at the KL divergence between both heuristics and BMI, which is now substantially smaller for the equal weighting heuristic as shown in Figure 4.3 (d).

We have seen that BMI discovered different heuristics in two classes of environments. Next, we show that there are also environments where this is not the case. For this, we optimized MI and BMI such that they adjust to problems without additional information in the form of ranking or direction. Gini coefficients obtained from BMI reveal that neither single cue nor equal weighting heuristics are resource-rational under such circumstances, as shown in Figure 4.4 (a). Instead, the pattern now looks more similar to one observed in MI and the ideal observer models, shown in Figures 4.4 (b) and (c) respectively. In all cases, Gini coefficients cover the full range of possible values, indicating that inferred weight vectors integrate information from multiple features to different degrees. This time, we find no difference in the KL divergence between both heuristics and BMI (ref. Figure 4.4 (d)), which confirms the earlier conclusion that BMI does not recover any of the two heuristics in an environment without additional information about ranking or direction.

4.4.4 Experimental Predictions

BMI discovers both single cue and equal weighting heuristics when information about ranking and direction is provided, respectively. However, resulting strategies diverge from known heuristics whenever such information is not present. Instead, our simulation results suggest that weighted combinations of multiple features should be used in such situations. Under the assumption that people make adaptive and computationally efficient inferences, our results enable us to make precise predictions about when to expect heuristics as part of human decision-making and when not: knowing the correct ranking of attributes leads to one reason decision-making, knowing the directions of the attributes leads to equal weighting, and not knowing about either leads to strategies that use weighted combinations of multiple attributes. Below, we present the results of three paired comparison studies that confirm the predictions made by BMI.

4.5 Experiment 1: Known Ranking

In the first study, participants made decisions in multiple paired comparison tasks while having access to a ranking of features, but not their underlying weights. Previously, we showed that in environments with known feature rankings, single cue heuristics are resource-rational strategies. Hence, we hypothesized that people are more likely to apply the single cue heuristic in this condition.

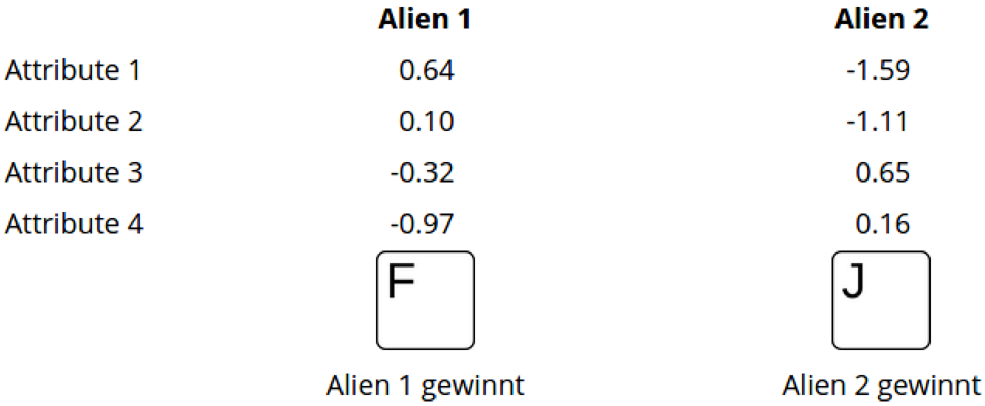
4.5.1 Methods

Participants

Participants were students from the University of Marburg, taking part in the study for course credits. Besides course credits, they got a chance to win a €10 voucher if they made more than 66.6% correct decisions. The experiment was approved by the local ethics board (AZ 2020-32k). In total, we collected data from 28 participants (23 female, average age: 22.36 ± 5.65).

Procedure

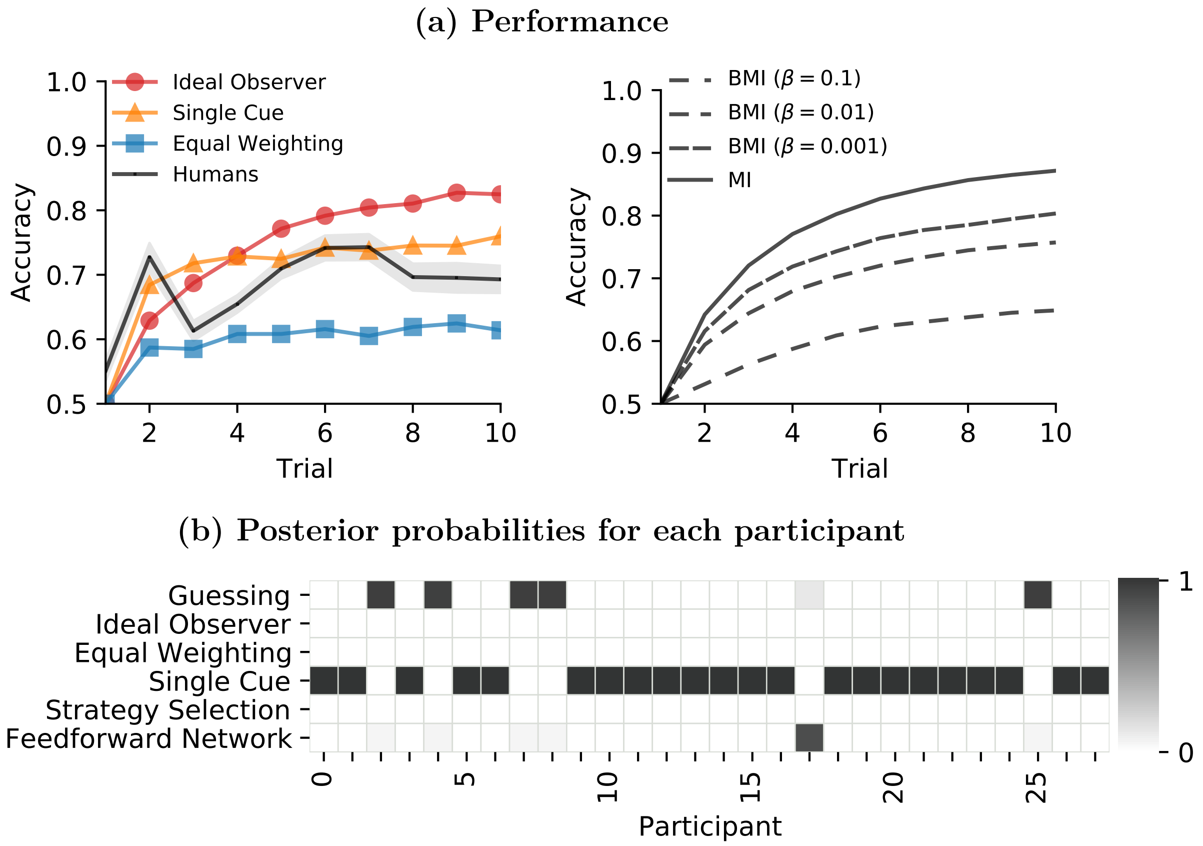
Each participant performed 30 different paired comparison tasks that were randomly generated according to the previously described distribution. Each task consisted of ten trials. Underlying weights remained fixed within a task but varied between tasks. Participants were informed about transitions between tasks. Each participant encountered the same set of paired comparison tasks in a randomized order.



	Alien 1	Alien 2
Attribute 1	0.64	-1.59
Attribute 2	0.10	-1.11
Attribute 3	-0.32	0.65
Attribute 4	-0.97	0.16
	F	J
	Alien 1 gewinnt	Alien 2 gewinnt

Figure 4.5: Graphical illustration of a single trial in the experiment. "Alien X gewinnt" translates to "Alien X wins".

The problem was framed as an alien sports competition on an unknown planet (see Figure 4.5). Participants observed four numerical attributes for two aliens and indicated by a button press which alien they believed was more likely to win. The alien cover story was used to keep the meaning of features completely abstract from the participant's perspective. Participants did not have access to the underlying weights but instead had to learn about the importance of features based on experience. Feedback about the correct choice was provided directly after each decision. For this condition, features were displayed in descending order based on the magnitude of weights. Participants were told that features are arranged from top to bottom according to how well they predict the winner. Being aware of this additional ranking information allowed them to apply strategies that are appropriate for this environment. Participants went through a short tutorial and did a comprehension check to confirm that they understood the instructions. The median time to complete the experiment was 26.00 minutes.



(a) Performance

(b) Posterior probabilities for each participant

Figure 4.6: (a) Percentage of correct decisions (averaged over all tasks) in the ranking condition plotted over the number of trials within a task. For human performance shaded contours represent the standard error. The left panel shows the ideal observer model and both heuristics, while the right shows BMI for different values of β. (b) Posterior distributions for each participant over different strategies in the ranking condition. High values indicate that the participant was likely to use the corresponding strategy.

4.5.2 Results

Performance

Figure 4.6 (a) shows the percentage of correct decisions for participants in our study together with the accuracy of different models. Participant performance was within the range of the single cue heuristic and BMI. On average, participants made $68.25 \pm 7.55\%$ correct choices.

Model Comparison

If people make efficient use of their available computational resources, we expect them to adopt the single cue heuristic in this experiment. To examine this hypothesis, we performed a Bayesian model comparison and computed posterior probabilities of different models given the decisions made by a participant. Appendix D provides a detailed description of the methods we used for statistical analysis. Because the single cue heuristic and BMI make redundant predictions, we decided to split our analysis into two parts. First, we analyzed all models except BMI for individual participants. Then, we compared BMI against the other models on the data of all participants.

In 22 out of 28 participants, we found evidence for the application of the single cue heuristic. For all of those participants, the model evidence decisively favored the single cue heuristic ($p(m = \text{SC}|\mathcal{D}_i) > 0.99$). Figure 4.6 (b) summarizes posterior probabilities of different models for all participants. Most of the participants not best described by one reason decision-making were instead best described by guessing; one participant was best described by the feedforward network. The *protected exceedance probability* (PXP), which measures the probability that a particular model is more frequent in the population than all the other models under consideration (Rigoux et al., 2014), favored the single cue heuristic decisively (PXP > 0.999).

Finally, we compared how well BMI fared against the other models. Because BMI also includes guessing with large and compensatory strategies with low resource limitations, it allows us to capture individual differences. The resulting posterior probabilities indicated that across all participants BMI offered an even better explanation for the observed data than the other models ($p(m = \text{BMI}|\mathcal{D}) \approx 1$). This is the case, because BMI explained the behavior of participants that used single cue heuristics *and* participants that used guessing.

4.5.3 Discussion

Most empirical evidence for one reason decision-making has been provided by studies that involved a cost for acquiring information about features (Bröder, 2000, Rieskamp and Otto, 2006, Bröder and Gaissmaier, 2007). However, even with an experimental protocol that favored a few pieces of information, evidence for these strategies remained inconclusive (Newell et al., 2003, Scheibehenne et al., 2013). When information is freely available, people are often better described through compensatory strategies such as logistic regression (Bröder, 2000, Lee and Cummins, 2004, Glöckner and Betsch, 2008, Parpart et al., 2018). Our results are among the first to decisively show that people's choices can be based on a single piece of information, even when such strategies are not favored by the experimental protocol. This was possible, because we precisely identified conditions under which one reason decision-making *should* appear. Nearly all participants in our study applied strategies that were efficient in terms of resources while also accounting for environmental characteristics.

4.6 Experiment 2: Known Direction

In our second study, we provided no information about ranking and instead informed participants about feature directions; otherwise, it was identical to the first experiment. In our previous analysis, we have seen that this modification also caused a change in what strategy is resource-rational. Now, resource-rational decision-making amounts to the application of equal weighting heuristics. We, therefore, hypothesized that participants would become more likely to use such strategies.

4.6.1 Methods

Participants

Participants were students from the University of Marburg, taking part in the study for course credits. Besides course credits, they got a chance to win a €10 voucher if they made more than 66.6% correct decisions. The experiment was approved by the local ethics board (AZ 2020-32k). In total, we collected data from 24 participants (22 female, average age: 22.54 ± 3.28).

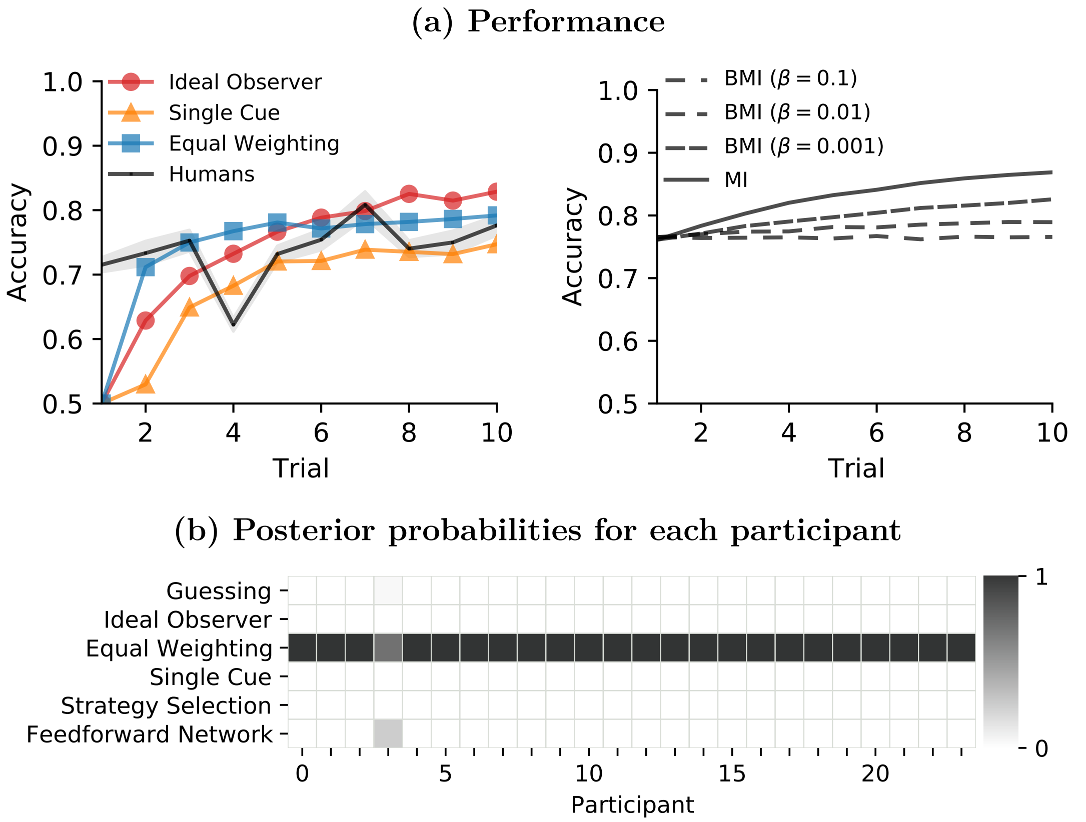
Procedure

The design was identical to the first experiment, except that participants were informed about the presence of positive feature directions instead of the feature ranking. This was realized by telling them that higher feature values always made it more probable for an alien to win the competition. The median time to complete the experiment was 29.69 minutes.

4.6.2 Results

Performance

Participants made on average 73.85 ± 4.53% correct choices, putting their performance within the range of all models, see Figure 4.7 (a). The higher average


(a) Performance

(b) Posterior probabilities for each participant

Figure 4.7: (a) Percentage of correct decisions (averaged over all tasks) in the direction condition plotted over the number of trials within a task. For human performance shaded contours represent the standard error. The left panel shows the ideal observer model and both heuristics, while the right shows BMI for different values of β. (b) Posterior distributions for each participant over different strategies in the direction condition. High values indicate that the participant was likely to use the corresponding strategy.

performance indicates that participants found it overall easier to process information about direction than about ranking. Participants' performance in the initial step turned out to be substantially higher than the ideal observer model and both heuristics, indicating that information about direction is useful even before making observations. This characteristic is also captured in BMI.

Model Comparison

In this condition, equal weighting and BMI made partially redundant predictions. Thus, we again decided to split our analysis into two parts. First, we analyzed all models except BMI for individual participants. Then, we compared BMI against the other models on the data of all participants.

The posterior probabilities of different models, illustrated in Figure 4.7 (b), confirmed the prediction of our earlier simulations. Most participants indeed adhered to the resource-rational maxim and applied equal weighting heuristics. For all participants, equal weighting provided the best explanation for the observed data. For all but one participant, evidence turned out to be decisive ($p(m = \text{EW}|\mathcal{D}_i) > 0.99$). The probability that equal weighting was the most frequent model in the population (PXP > 0.999) supported the conclusion that people, in general, applied equal weighting heuristics when information about direction was available.

When additionally comparing BMI against the other models on the aggregated data of all participants, we found that BMI again offered an even better explanation than all other models ($p(m = \text{BMI}|\mathcal{D}) \approx 1$). Here, this was the case because BMI was able to capture participants' decisions in the initial step, while the equal weighting heuristic did not.

4.6.3 Discussion

Similar to the results of our first study, we found that people apply resource-rational strategies that are adequate for the given environment. Participants performed better compared to the first study, indicating that they found it easier to work with directions than with rankings. We speculate that one explanation for this observation could be that positive correlations are more frequently encountered in the world.

Previous empirical studies (see our earlier analysis in Tables 4.1 and 4.2) on heuristics were often restricted to tasks with positive correlations between features and the criterion. Despite this, few studies actually consider equal weighting heuristics when comparing their hypotheses. Instead, most of them attempted to show that people rely on one reason decision-making, often with inconclusive results. We believe that this mismatch between the hypotheses being tested and the structure of the tasks considered is an important factor in explaining the mixed results of prior empirical work.

4.7 Experiment 3: Unknown Ranking and Direction

In our final study, we investigated choice behavior in an environment that did not provide information about ranking or direction. In the previous model simulations, we have demonstrated that no heuristic emerges under such conditions. Instead, BMI discovered strategies with compensatory weights even under large resource constraints. Hence, we predicted that people in this condition are less reliant on traditional heuristics and instead integrate information from multiple features properly.

4.7.1 Methods

Participants

Participants were students from the University of Marburg, taking part in the study for course credits. Besides course credits, they got a chance to win a €10 voucher if they made more than 60% correct decisions. The experiment was approved by the local ethics board (AZ 2020-32k). In total we collected data from 23 participants (16 female, average age: 23.09 ± 4.38).

Procedure

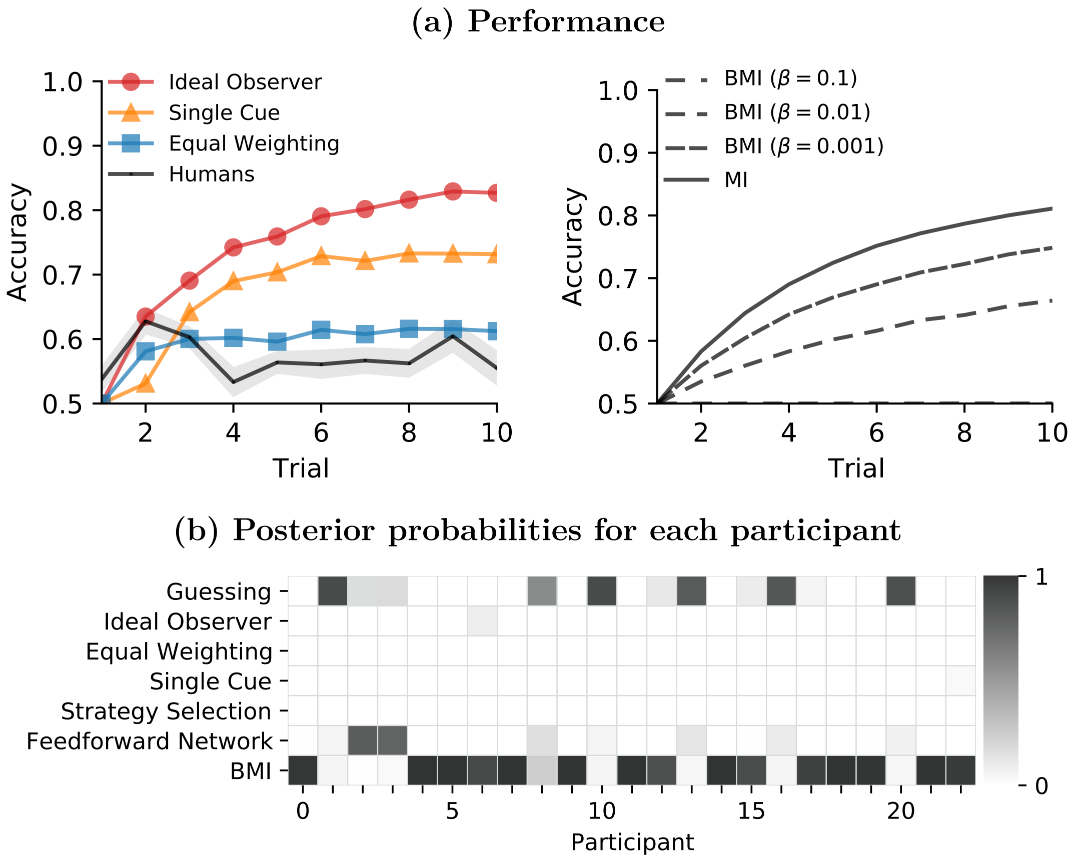
The design was identical to both previous experiments, except that it did not include information about feature rankings and direction anymore. The median time to complete the experiment was 36.09 minutes.

4.7.2 Results

Performance

The ideal observer model and the equal weighting heuristic remained identical in their performance compared to the first study, see Figure 4.8 (a). The single cue heuristic however performed slightly worse, as it was not provided with knowledge about the most predictive feature anymore, but instead had to infer it based on observations. Note, that with an identical level of resource limitations the performance of BMI substantially decreased compared to the previous environments.

Participants also found this version much harder and performed substantially worse. Without the additional information from the first two conditions, their


(a) Performance

(b) Posterior probabilities for each participant

Figure 4.8: (a) Percentage of correct decisions (averaged over all tasks) in the unrestricted condition plotted over the number of trials within a task. For human performance shaded contours represent the standard error. The left panel shows the ideal observer model and both heuristics, while the right shows BMI for different values of β. (b) Posterior distributions for each participant over different strategies in the unrestricted condition. High values indicate that the participant was likely to use the corresponding strategy.

cognitive resource limitations became a dominating factor. The average performance dropped to 57.14 ± 4.38%. While some participants performed well, a substantial amount was at or close to chance level.

Model Comparison

According to our model simulations, we should expect to find evidence for models using weighted combinations of multiple features in this condition. Because no known heuristic emerged in this environment, we did not split our analysis and already considered BMI on the level of individual participants.

Posterior probabilities obtained from a Bayesian model comparison in Figure 4.8 (b) confirmed that most participants combined information from multiple features instead of using heuristics like equal weighting or one reason decision-making. Fifteen out of 23 participants were best described by BMI; in eight of those we found decisive evidence ($p(m = \text{BMI}|\mathcal{D}_i) > 0.99$). Amongst the participants not best described by BMI, six were best described by guessing and two by the feedforward network. We again found that BMI fared favorably against all other models on the aggregated data ($p(m = \text{BMI}|\mathcal{D}) \approx 1$). The protected exceedance probability (PXP > 0.999) also supported the conclusion that BMI was the most frequent explanation for participants in our population.

4.7.3 Discussion

In an environment that did not provide additional information about ranking or direction, participants' decision-making again followed the prediction made by BMI. Most participants applied strategies that involved weighted combinations of features, as it was suggested by our model simulations. The general result that most people were able to quickly combine information from multiple sources if needed is also consistent with results of prior studies (Bröder, 2000, Glöckner and Betsch, 2008, Parpart et al., 2018).

4.8 General Discussion

At the core of theories of ecological rationality, researchers have posited an interaction between cognition and the environment. Brunswik (1956) argued that human perception cannot be understood in laboratory settings alone, but rather has to be interpreted in the light of real environments in which real objects are perceived and acted upon. Simon (1990b) famously highlighted the interaction between cognition and the environment using an analogy of a pair of scissors, with one blade being the structure of the environment and the other blade the computational capabilities of the subject. This conceptualization of ecological rationality has strongly influenced theories of heuristic decision-making. The need to economize cognitive resources places pressure on the mind to employ heuristics that work well in specific environments. Nonetheless, how people pick a particular heuristic for a specific environment and where those heuristics come from in the first place has remained elusive. The theoretical picture becomes even more puzzling when looking at the empirical support for heuristic decision-making. Proponents of heuristic decision-making acknowledge these problems. For example, Gigerenzer (2008) writes: "Why do heuristics work? They exploit evolved capacities that come for free. In addition, they are tools that have been customized to solve diverse problems. By understanding the ecological rationality of a heuristic, we can predict when it fails and succeeds. The systematic study of the environments in which heuristics work is a fascinating topic and is still in its infancy." But what does a theory, which can explain how heuristics emerge and how they are selected while at the same time accounting for the sometimes mixed empirical results, look like?

We have put forward BMI as a theory that makes significant advances on these questions. Our simulation results show that BMI discovers previously suggested heuristics. Thus, it provides a normative justification for heuristic decision-making. Moreover, we find that different heuristics emerge depending on environmental assumptions. Thus, BMI also explains how decision-making strategies are selected. Finally, our account generates predictions about if and when a specific heuristic should be applied. Since we find that one reason decision-making is unlikely to occur in many of the past experimental set-ups, this explains the mixed results of prior empirical work.

Already early on, researchers working on heuristic decision-making levied the criticism that simply observing behavioral biases is not enough, and that "in place of plausible heuristics that explain everything and nothing – not even the conditions that trigger one heuristic rather than another – we need models that make surprising (and falsifiable) predictions" (Gigerenzer, 1996). However, the very fact that several heuristic components have been claimed to be part of a heuristic toolbox without fully specifying how they are selected and combined, has subjected heuristic theories to a similar line of criticism: "... if one cannot predict which heuristics will be used in which environments then determining the heuristic that will be selected from the toolbox for a particular environment becomes necessarily post hoc and thus the fast-and-frugal approach looks dangerously like becoming unfalsifiable." (Newell et al., 2003). In contrast to these arguments, BMI makes clear, falsifiable and surprising predictions about when people should apply which heuristic. Specifically, our simulation results show that there are three important classes of environments triggering three decision-making strategies. If people know the correct ranking of attributes but not their weights, then they should exhibit one reason decision-making. If people know the direction of the attributes but not their ranking, then they should exhibit equal weighting strategies. Finally, if people do not know either the ranking or the direction of the attributes, then they should exhibit strategies that use weighted combinations of attributes.

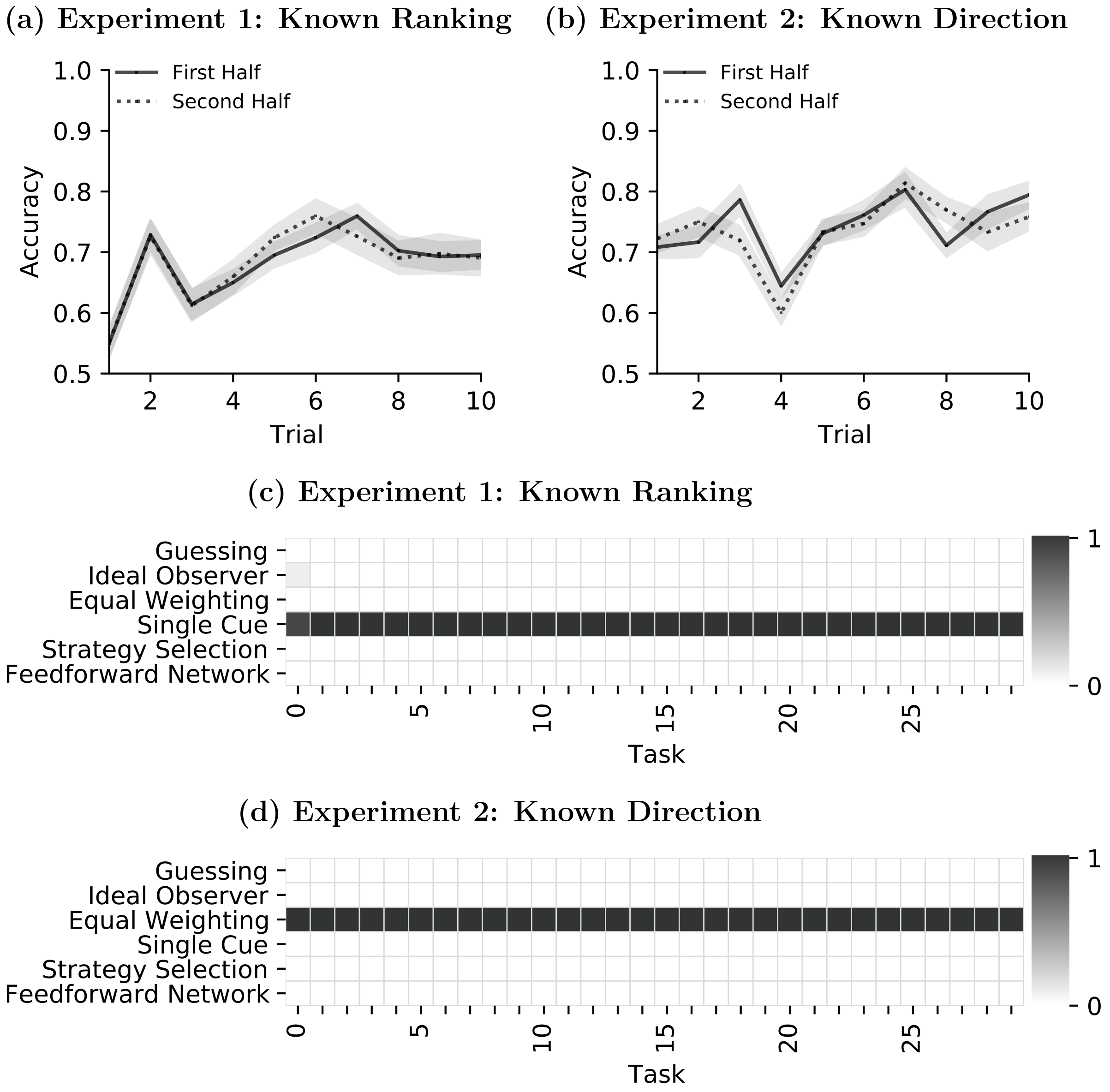
We subjected these predictions to a rigorous test in three paired comparison experiments and found that the vast majority of participants applied decision-making strategies as predicted by BMI. Moreover, BMI captured elements of human decision-making that could not be explained by traditional heuristics in all three experiments: In the first study it additionally accounted for participants that resorted to guessing, in the second study it provided an explanation for the good initial performance of participants and in the third study it predicted correctly that performance should decrease and that people apply compensatory strategies instead of established heuristics. Together, these results enrich our theoretical and empirical understanding of ecologically rational decision-making.

4.8.1 Limitations

Gigerenzer and Todd (1999) argued that decision-making under limited resource cannot be expressed through models that perform optimization under constraints: "Optimization under constraints also limits search, but does so by computing the optimal stopping point, that is, when the costs of further search exceed the benefits." Computing this optimal stopping point can be at least as expensive as finding the optimal solution; hence it defeats the initial intention of modeling decision-making under resource limitations (Gigerenzer and Todd, 1999, Scheibehenne and Von Helversen, 2009). BMI involves optimization under constraints but importantly does so at the meta-learning level, which happens on a much larger time scale (e.g. through evolutionary processes). Learning within an individual task, on the other hand, is fast as it does not involve any form of optimization. This perspective of learning at multiple scales is also at the core of recent theories of fast and slow reinforcement learning (Botvinick et al., 2019).

BMI assumes that meta-learning happened prior to the experiment, but it remains agnostic about the exact processes controlling the acquisition of strategies. BMI could, for example, be acquired through evolutionary processes, through individual experiences, or both. If meta-learning indeed happened prior to the experiment, we should find no noticeable improvement in performance over the course of our studies. We find support for this hypothesis when comparing human performance in the first and second half of our studies (Figure 4.9 (a) and (b)). Furthermore, we evaluated posterior probabilities of different models for each task as opposed to for each individual participant (Figure 4.9 (c) and (d)) and found that participants did not apply different strategies during the experiment. Nonetheless, a valid criticism of our current work is that it does not address the precise process of meta-learning, and whether this process is rather shaped by ontogeny, phylogeny, or both. This is indeed an open problem for all theories of heuristic decision-making, which at various times have argued that heuristics emerge from evolutionary pressures (Hutchinson and Gigerenzer, 2005), developmental processes (Gigerenzer, 2003), or task-specific adaptations (Marewski and Schooler, 2011). The time scale of meta-learning therefore remains an open theoretical and empirical question.

Currently, our approach also does not directly offer a way to predict which properties of the environment will determine what type of decision-making strate-


(a) Experiment 1: Known Ranking

(b) Experiment 2: Known Direction

(c) Experiment 1: Known Ranking

(d) Experiment 2: Known Direction

Figure 4.9: (a) and (b) show that performance of participants did not change over the experiment, indicating that meta-learning already happened prior to the experiment. Shaded contours represent the standard error. (c) and (d) confirm this observation by showing that the selection of strategies also did not change during the experiment. High values indicate that the corresponding strategy was applied with high probability in the given task.

gies are ecologically rational. Instead, we have to train our meta-learning models in different environments and then analyze what decision strategies emerge, for example by analyzing the weights' Gini coefficient. Looking at a model's emerging properties is a common method when neural network approaches are applied to psychological questions (Ritter et al., 2017). We believe that this possible weakness can also be a strength, because it forces researchers to truly study the properties of environments, as has been the core proposal of theories of ecological rationality for decades.

4.8.2 Related Work

To highlight what BMI adds to existing theories, we compare it to other ideas put forward in previous investigations. In the context of decision-making, we focus on methods that address how strategies are selected and how they are discovered. Beyond that, we discuss how meta-learning and resource-rationality have been applied to understand other phenomena.

Strategy Selection

First, there have been several theories explaining how strategies are selected. Rieskamp and Otto (2006) proposed a theory of strategy selection learning that framed the strategy selection process as a model-free reinforcement learning problem. Their theory assumes that people slowly learn how to select the right strategy from a given repertoire of strategies based on repeated interactions. A key finding of their experiments was that over time participants learned to select the best-performing strategy for a particular environment. Their method requires learning from scratch whenever it encounters novel problems and hence it does not address how knowledge is transferred between different environments, or why participants are immediately able to select appropriate strategies in our experiments.

Lieder and Griffiths (2017) addressed the missing ability to transfer knowledge between environments through an approach based on *rational meta-reasoning.* Based on properties of the environment, they predicted speed and accuracy of different strategies. They showed that participants selected the strategy that was best for solving the speed-accuracy trade-off in the current context. In contrast to their work, we used separate models for each environment. However, it would be possible to extend our modeling framework by conditioning the initial state of the RNN on features of an environment.

Marewski and Schooler (2011) postulated a probability landscape describing an individual's ability to apply a strategy as a function of cognitive capabilities and the environment. Their work referred to situations in which a strategy can be applied as a *cognitive niche* and showed that cognitive niches of different strategies are disjoint in many cases. This greatly simplified the strategy selection problem and was in line with participants' behavior across a number of experiments. We believe that cognitive niches could also be the result of meta-learning, where an algorithm adapts to a given characteristic of an environment until it cannot easily be applied to a vastly different environment anymore.

Previous theories of strategy selection all require to define a set of potential strategies in advance. In contrast, BMI is not restricted to predefined sets and instead discovers useful strategies on the fly.

Strategy Discovery

There have also been some accounts that explain how strategies are discovered. Schulz et al. (2016a) proposed a method for learning decision-making strategies from small, probabilistic building blocks. Based on a self-reinforcing sampling scheme, they were able to build tree-like non-compensatory heuristics. Their approach can recover TTB on data sets that have been generated by the TTB heuristic. However, it is not able to learn about other, non-compensatory strategies or to make predictions about when participants would prefer which strategy.

Lieder et al. (2017) suggested a model that composes strategies from atomic computations. According to their theory, an agent represents computations as costly actions in a meta-level Markov decision process. The agent's goal is to maximize the external payoff obtained from making correct decisions while accounting for the computational costs of actions. When they applied their theory to several decision-making problems, they found that it discovered two known heuristics – TTB and guessing – as well as a novel strategy that combined TTB with satisficing (Simon, 1956).

Parpart et al. (2018) showed that heuristics can emerge from Bayesian inference in the limit of infinitely strong priors. Using this idea, they identified priors corresponding to an equal weighting heuristic. Finding a prior that leads to TTB proved to be more challenging in the Bayesian framework and was only possible after introducing an additional decision rule. Instead of relying on the complexity argument as justification for heuristics, their analysis suggested that heuristics work well because they implement priors that reflect the actual structure of the environment.

Theories that build algorithms from simpler computations (Schulz et al., 2016a, Lieder et al., 2017) discover one reason decision-making heuristics without difficulties, but struggle to account for equal weighting heuristics. Theories based on Bayesian inference (Parpart et al., 2018) on the other hand have no difficulties with discovering equal weighting heuristics, but require additional components to find heuristics that rely on a single piece of information. We show that people actually use both classes of strategies and provide a theory that can discover both of them in an appropriate context. While there exist prior approaches that address either the strategy selection problem or the strategy discovery problem independently, BMI is also the first to account for both problems jointly within a unified framework.

Meta-Learning as Theory of Human Behavior

Brighton (2006) and Chater et al. (2003) considered standard feed-forward networks trained with backpropagation as models of decision-making in paired comparison tasks. Their results indicated that, if only a few examples were used, such models tended to overfit and were outperformed by much simpler, more robust alternatives. Brighton (2006) suggested meta-learning as a potential solution to this problem of overfitting but did not provide a concrete implementation of this conjecture. BMI is such an implementation that can be applied to paired comparison tasks with few examples and – crucially – *without* showing signs of overfitting. The key to BMI's success is that learning happens in the network's recurrent activations and not through traditional gradient-based training schemes.

When we look beyond decision-making and paired comparison tasks, meta-learning has recently received increased attention as an explanation for human behavior across a variety of cognitive and neuroscientific questions. For example, meta-learning has been shown to lead to human-like characteristics in the contexts of few-shot learning (Santoro et al., 2016), systematic compositionality

(Lake, 2019), exploration (Binz and Endres, 2019) as well as one-shot navigation and model-based reasoning (Wang et al., 2016). The current work adds heuristic strategies of decision-making as another domain to this list.

Directly relevant to our work is the approach of Dasgupta et al. (2020), who taught neural networks to approximate Bayesian inference, given some information about an inference problem's prior and likelihood. Restricting the size of the network allowed them to account for a large number of cognitive biases, including base rate neglect and conservatism. This approach shares its core principles with our theory: resource-rationality and meta-learning. However, BMI does not approximate Bayesian inference explicitly as done by Dasgupta et al. (2020). Instead, it attempts to infer distributions that are optimal for making future predictions (which may or may not correspond to Bayesian inference).

4.8.3 Future Directions

Most computational models in psychology and cognitive science are confined to idealized settings. BMI on the other hand can – in principle – scale to much more complex domains (Wang et al., 2016, Santoro et al., 2016). Having access to such models allows us to study human behavior under more realistic conditions. In the context of decision-making, it becomes, for example, possible to investigate how and why different representational formats influence human strategies (Bröder and Schiffer, 2006) by learning models that directly process visual representations of the task.

In this paper, we have applied BMI to the paired comparison setting. However, BMI is more general than that and we believe that it could also be used to explain heuristics in other contexts, such as the recognition heuristic (Goldstein and Gigerenzer, 2002) or the gaze heuristic (Shaffer et al., 2004, Belousov et al., 2016). BMI could also provide insights into other phenomena in human learning, such as the observation that learning about multiple tasks is usually easier when tasks are presented successively compared to an interleaved presentation (Flesch et al., 2018).

The classical approach to computational modeling is to propose a model, test its predictions, and finally revise the model if required. However, we can also envision an approach for the revision of theories that puts the study of environments first. In this framework, we would ask ourselves what environments can account for observed behavior assuming that people make ecologically and resource-rational decisions, instead of revising arbitrary parts of the model. That this is a promising research direction for building more human-like agents was shown for example by Hill et al. (2020), who demonstrated that systematic generalization can be an emergent property of an agent interacting with a *rich* environment.

Finally, our theory provides us with a set of predictions about what should happen when available computational resources are manipulated. It will be interesting to see whether people follow the behavioral trajectories stipulated by BMI when put under cognitive load or whether patients with attention or memory impairment are better described by models with lower complexity.

4.8.4 Conclusion

The idea that theories of human cognition should consider both the structure of the environment and the computational capabilities of the subject has been a central theme in psychology (Simon, 1990b, Todd and Gigerenzer, 2012). However, actual implementations of this principle have been lacking so far. BMI provides such an implementation by combining the ideas of resource-rationality and meta-learning. BMI accounts for two open questions in the decision-making literature simultaneously, explaining why different strategies emerge and how appropriate strategies are selected. By mapping out environments that cause different strategies to be resource-rational, we obtain precise predictions about when previously suggested heuristics should be used and when not. We confirmed these predictions in three paired comparison experiments. Taken together, BMI offers a normative and empirically supported theory of human decision-making.

5

Towards A Domain-General Theory Of Order Effects

Abstract: Order effects are ubiquitous in human learning: people's responses vary when learning about associations of events if the order of trials is rearranged, they learn faster about functions with a structured presentation of datapoints, and they adapt better to multiple tasks simultaneously when encountering them in blocks. We show that while previous theories reproduce order effects found in associative learning studies, they cannot provide a unifying explanation for the human sensitivity to rearrangements of observations in other domains, such as function learning and multi-task learning. To close this explanatory gap, we suggest BMI as an alternative theory for why order effects occur. Through model simulations, we show that BMI captures all order effects under consideration without the need for any domain-specific modifications. These results offer a significant step towards a domain-general computational theory of order effects.

This chapter is based on the following publication:

Binz, M., & Endres, D. (2020). Towards A Domain-General Theory Of Order Effects (submitted).

5.1 Introduction

Humans are sensitive to the arrangement of data-points during learning. Empirically, such order effects have been observed consistently across different subdisciplines of cognitive psychology: in associative learning multiple paradigms in which people's responses vary when the order of trials is rearranged have been identified (Been et al., 2003, Medin and Bettger, 1991), in function learning people find it easier to learn about functions when inputs are presented in ascending order (Byun, 1996) and when learning about multiple tasks simultaneously human performance can improve if tasks are presented in blocks compared to an interleaved presentation (Lee et al., 1992, Flesch et al., 2018).

This behavior is in stark contrast to many machine learning models that assume that data-points are independent and identically distributed, or at least exchangeable, which in turn implies that learning in such models is invariant to a reordering of data-points. This is true in particular for Bayesian theories of learning, which in general have been very successful at capturing how people learn (Anderson, 1991a, Dayan and Long, 1998, Gershman, 2015, Lucas et al., 2015), but often assume that data-points are exchangeable. Previous research has suggested two approaches to resolve this dilemma while allowing to retain the normative character of Bayesian inference:

1. People assume that their environment is changing over time (Dayan and Long, 1998, Dayan and Kakade, 2001, Courville et al., 2006).

2. People only use approximations to exact Bayesian inference (Kruschke, 2006, Daw et al., 2008, Sanborn et al., 2010, Abbott and Griffiths, 2011,

Sanborn and Silva, 2013).

A changing environment can manifest itself in order effects because under such conditions recent data provides more information about the environment and hence should be weighed more heavily than old data. Approximate inference strategies on the other hand inevitably discard some information about the data. Importantly, trial order can influence what information gets thrown away, which in turn induces order effects of different kinds. Both ideas are appealing from a psychological perspective. Realistic environments are often non-stationary, which makes it reasonable to assume that people carry the strategies acquired under such conditions over to laboratory studies. At the same time, people do not have unbounded computational resources (Simon, 1990a, Gershman et al., 2015, Lieder and Griffiths, 2020), and thus it is natural to consider approximate inference strategies that are less demanding in terms of their computational complexity.

We ask: how well do the current explanations for order effects generalize across domains, and what are the ingredients required for a domain-general theory of order effects? We find that while both previously described ideas capture order effects found in associative learning studies (Dayan and Kakade, 2001, Kruschke, 2006, Daw et al., 2008, Sanborn and Silva, 2013), they fall short of an explanation for order effects in other domains, such as function and multi-task learning.

This lack of a comprehensive account calls for a novel theory. In particular, we suggest that order effects arise from a meta-learned learning algorithm that optimally trades-off performance for a lower computational complexity. To test this hypothesis, we apply BMI to several paradigms in which people show order effects. BMI unifies – and generalizes – previous theories of order effects. It has been adapted to particular environment through meta-learning, allowing it to attribute order effects to properties of the environment. It furthermore involves a regularizer for the complexity of the emerging learning algorithm, allowing it to attribute order effects to limited computational resources.

The remainder of this chapter is structured as follows: we first provide background information about different models of human learning. Then, we look at three research areas where order effects occur: associative learning, function learning and multi-task learning. First, we verify that BMI reproduces order effects found in associative learning, which are also explained by earlier theories. Then, we show that BMI additionally captures order effects found in function and multi-task learning, which are challenging for alternative theories. Finally, we summarize our results and discuss their implications for both cognitive psychology and machine learning.

5.2 Computational Models

In this chapter, we consider regression problems, i.e., tasks in which an agent has to learn how to map an input variable $\mathbf{x} \in \mathbb{R}^d$ to a target variable $y \in \mathbb{R}$. Many experimental paradigms can be framed in this way. For example, in an associative learning experiment $\mathbf{x}$ may correspond to the presence or absence of different stimuli and y to an associated reward; or in a function learning setting $\mathbf{x}$ could correspond to different features of a tree, while y represents the tree's size in one month from now. In each task the agent encounters a sequence of input-target pairs $\mathbf{x}_{1:T}, y_{1:T}$. In each time-step t, it first observes an input

$\mathbf{x}_t$, then makes a prediction for that input, and subsequently receives feedback about the actual target variable y_t. In this section, we provide background information on different computational models that are sensitive to the order in which data-points are observed.

5.2.1 Kalman Filters

Kalman filters implement the idea of Bayesian inference for a specific class of linear dynamical systems. In particular, they make the following assumptions about prior and likelihood:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}_{0:T-1}) = \prod_{t=1}^{T} \mathcal{N}(y_t|\mathbf{w}_{t-1}^T \mathbf{x}_t, \sigma^2) \tag{5.1}$$

$$p(\mathbf{w}_{0:T-1}) = \mathcal{N}(\mathbf{w}_0|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{t=1}^{T-1} \mathcal{N}(\mathbf{w}_t|\mathbf{w}_{t-1}, \tau^2 \mathbf{I}) \tag{5.2}$$

What is the intuition behind these assumptions? Equation 5.1 states that the target variable can be represented through a linear combination between inputs and a set of regression weights plus some additive normally distributed noise with variance σ^2. Equation 5.2 expresses the additional assumption that parameters are jittered by normally distributed noise with variance τ^2 in between observations. In our upcoming model simulations, we set the initial prior to a standard normal distribution. Because all random variables are normally distributed, the posterior distribution in a Kalman filter has an analytical expression:

$$\begin{aligned} p(\mathbf{w}_t|\mathbf{x}_{1:t}, y_{1:t}) &= \mathcal{N}(\mathbf{w}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) && (5.3) \\ \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \mathbf{k}_t \left(y_t - \boldsymbol{\mu}_{t-1}^T \mathbf{x}_t\right) && (5.4) \\ \boldsymbol{\Sigma}_t &= \boldsymbol{\Sigma}_{t-1} + \tau^2 \mathbf{I} - \mathbf{k}_t \mathbf{x}_t^T \left(\boldsymbol{\Sigma}_{t-1} + \tau^2 \mathbf{I}\right) && (5.5) \\ \mathbf{k}_t &= \frac{\left(\boldsymbol{\Sigma}_{t-1} + \tau^2 \mathbf{I}\right) \mathbf{x}_t}{\mathbf{x}_t^T \left(\boldsymbol{\Sigma}_{t-1} + \tau^2 \mathbf{I}\right) \mathbf{x}_t + \sigma^2} && (5.6) \end{aligned}$$

and the predictive posterior distribution can also be expressed analytically:

$$p(y|\mathbf{x}, \mathbf{x}_{1:t}, y_{1:t}) = \mathcal{N}(y|\boldsymbol{\mu}_t^T \mathbf{x}, \mathbf{x}^T \boldsymbol{\Sigma}_t \mathbf{x} + \sigma^2) \qquad (5.7)$$

For $\tau = 0$, which we also refer to as stationary Kalman filter, the underlying system does not change over time. In this case, data-points are independent and identically distributed, making the model invariant to rearrangement of data-points. However, when $\tau > 0$ Kalman filters consider recent data-points as more important because these provide more information about the current state of the system, which in turn leads to a recency bias (Dayan and Kakade, 2001, Daw et al., 2008).

5.2.2 Variational Linear Regression

Exact Bayesian inference can be challenging from a computational perspective. Thus, it makes sense to entertain the possibility that human learning is based on approximations. The prime examples of such approximations are variational inference (Jordan et al., 1999) and sample-based methods (Geman and Geman, 1984). Prior work has shown that both approaches can lead to order effects similar to the ones found in human learning (Daw et al., 2008, Sanborn et al., 2010, Sanborn and Silva, 2013, Abbott and Griffiths, 2011).

Here, we focus on a variational approximation to the stationary Kalman filter, which is also know as *variational linear regression*. In variational inference, the problem of inference is phrased as an optimization problem: the true posterior is approximated through a family of parametrized distributions – in our case a multivariate normal distribution $q(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ – and the goal is to find the member in the approximating family that minimizes the KL divergence to the true posterior. This can be achieved through maximizing the evidence lower bound:

$$\mathcal{L}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) = \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}\left[\log p(y_t|\mathbf{x}_t, \mathbf{w})\right] - \mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)||q(\mathbf{w}|\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\right] \quad (5.8)$$

where the initial approximate posterior $q(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ is set to the prior. We refer the reader to Appendix B for further details on how Equation 5.8 is optimized.

This variational approach is equivalent to Bayesian inference whenever the true posterior is within the considered approximating family. However, from a computational perspective, it is often convenient to restrict the approximating family to simpler distributions. A common constraint – which we also adopt in our model simulations – is to restrict the covariance matrix $\boldsymbol{\Sigma}_t$ to be low-rank. Necessarily, there will be a loss of information if we make such simplifications. This causes a sensitivity to the arrangement of data-points even if the exact solution is order-invariant because what information is lost may depend on trial order.

5.2.3 Bounded Meta-Learned Inference

We also investigated whether order effects result from a meta-learned learning algorithm that optimally trades-off performance for a shorter description length. To test this hypothesis, we train a RNN to act as a learning algorithm using the framework presented in Section 3.1. In time-step $t + 1$ the network processes the current input vector $\mathbf{x}_{t+1}$ and the target from the previous time-step y_t. Through its recurrent activations, it has access to the entire history of previously observed input-target examples. Its outputs correspond to the mean μ_{t+1} and standard deviation σ_{t+1} of a normal distribution. Together, these parametrize the predictive posterior distribution $p(y_{t+1}|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \boldsymbol{\Theta})$. Figure 3.1 illustrates how the network processes a sequence of observations.

Initially, the RNN implements a random mapping. During meta-learning it is then turned into a resource-rational learning algorithm. This is accomplished by minimizing the BMI objective until convergence:

$$\begin{aligned}\mathcal{L}_{\text{BMI}}(\boldsymbol{\Lambda}) = \mathbb{E}_{q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})}\left[\mathbb{E}_{p(\mathbf{x}_{1:T}, y_{1:T})}\left[\sum_{t=0}^{T-1} - \log p(y_{t+1}|\mathbf{x}_{t+1}, \mathbf{x}_{1:t}, y_{1:t}, \boldsymbol{\Theta})\right]\right] \\ + \beta \text{KL}\left[q(\boldsymbol{\Theta}|\boldsymbol{\Lambda})||p(\boldsymbol{\Theta})\right] \qquad (5.9)\end{aligned}$$

After meta-learning is completed, the RNN acts as an environment-specific learning algorithm that makes optimal use of limited computational resources. In Appendix C we provide a full specification on the network architecture, meta-learning procedure and choice of prior.

There are two reasons why BMI can exhibit order effects. First, BMI has been trained to make optimal inferences on a particular distribution over tasks. Thus, like Kalman filters, BMI can show order effects if the arrangement of data-points is important for reasoning accurately in a particular environment. However, unlike Kalman filters, BMI is not limited to linear environments that change according to normally distributed noise. Second, BMI embodies the idea that computational resources are costly. Thus, like variational inference, it can show order effects even if the training distribution implies that data-points are exchangeable. However, unlike variational inference, BMI does not require to make any assumptions about what hypotheses to consider in advance but instead adapts automatically to the demands of the environment during meta-learning.

5.2.4 Model Summary

Thus far, we have described three different models that can show sensitivities to the arrangement of observations: Kalman filters, variational linear regression, and BMI. Let us briefly recapitulate and contrast these approaches before moving to our model simulation results.

Kalman filters Explains order effects through optimal learning in a non-stationary environment. In a changing world, recent observations provide more information about the state of the world than past observations, which implies that a rational learner should exhibit a recency bias.

Variational linear regression Explains order effects through optimal learning within a simplified space of hypotheses. Simplifying the space of possible posterior distributions leads to a loss of information, which in turn may cause sensitivities to the arrangement of observations.

BMI Explains order effects through a meta-learned learning algorithm that optimally trades-off performance and complexity. Therefore, it provides two reasons for the emergence of order effects: (1) adaptation to the environment and (2) optimal use of limited computational resources.

5.3 Model Simulations

Next, we investigate how far towards a domain-general theory of order effects each of the aforementioned approaches brings us. For this, we look at three different areas where order effects occur: associative learning, function learning, multi-task learning. Previous work applied a combination of Kalman filters and approximate inference to capture different order effects found in associative learning studies (Daw et al., 2008). First, we reproduce these modeling results and additionally demonstrate that BMI also captures these effects. Then, we ask: how well does each theory generalize to other domains? We find that both Kalman filters and variational linear regression do not readily account for order effects from the function and multi-task learning literature, whereas BMI does.

5.3.1 Associative Learning

Associative learning studies how people learn associations between different events. We know from prior work that the order in which stimuli are presented influences how people learn such associations (Shanks, 1985, Kruschke, 2003). Here, we consider two widely studied paradigms from the associative learning literature in which people exhibit trial order effects: forward/backward blocking and highlighting.

Most associative learning studies are concerned with the response to a specific sequence of data-points and remain vague about the distribution of tasks an agent can potentially encounter. However, for our meta-learning models it is necessary to specify this distribution over tasks. Here, we make the following assumptions:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \prod_{t=1}^{T} \mathcal{N}(y_t|\mathbf{w}^T\mathbf{x}_t, 0.1) \tag{5.10}$$

$$p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_t|\mathbf{0}, \mathbf{I}) \tag{5.11}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) \tag{5.12}$$

Note that this training distribution is equivalent to the data generating distribution assumed by the stationary Kalman filter, implying that MI will be invariant to rearrangements of data-points because it simulates a stationary Kalman filter. For the associative learning simulations we encode different stimuli through binary vectors that indicate which stimulus is present (e.g. $A = [1, 0, 0]$ and $AB = [1, 1, 0]$) and train models on sequences of length 20.

Forward/Backward Blocking

In forward blocking stimulus A is paired with reinforcement Y several times during a first phase. In a second phase, the compound stimulus AB is paired with Y an equal number of times. Backward blocking is identical to forward blocking, except that the first and second phase are reversed. We use four observations per phase in our simulations and set Y to a constant value of 1. The forward and backward blocking paradigms are illustrated graphically in Figure 5.1 (a) and (b). Their ordering aside, data-points in forward and backward blocking are identical. Thus, models that make the assumption of exchangeability predict the same response to any stimulus at the end of the second phase for both paradigms. However, people typically show a weaker response to B after forward blocking compared to after backward blocking (Shanks, 1985).

Looking at Figure 5.1 (c), we observe that all models except MI can replicate the empirically observed effect. The observation that MI shows no difference between both paradigms comes as no surprise, as it approximately simulates a stationary Kalman filter, which is invariant under a rearrangement of data-points. All other models respond weaker to B after forward blocking compared to after backward blocking. Hence, there are at least three possible explanations for the order effect found in forward/backward blocking:

- Participants assume a non-stationary environment.
- Participants make inferences within a simplified space of hypotheses.
- Participants make optimal use of limited computational resources.

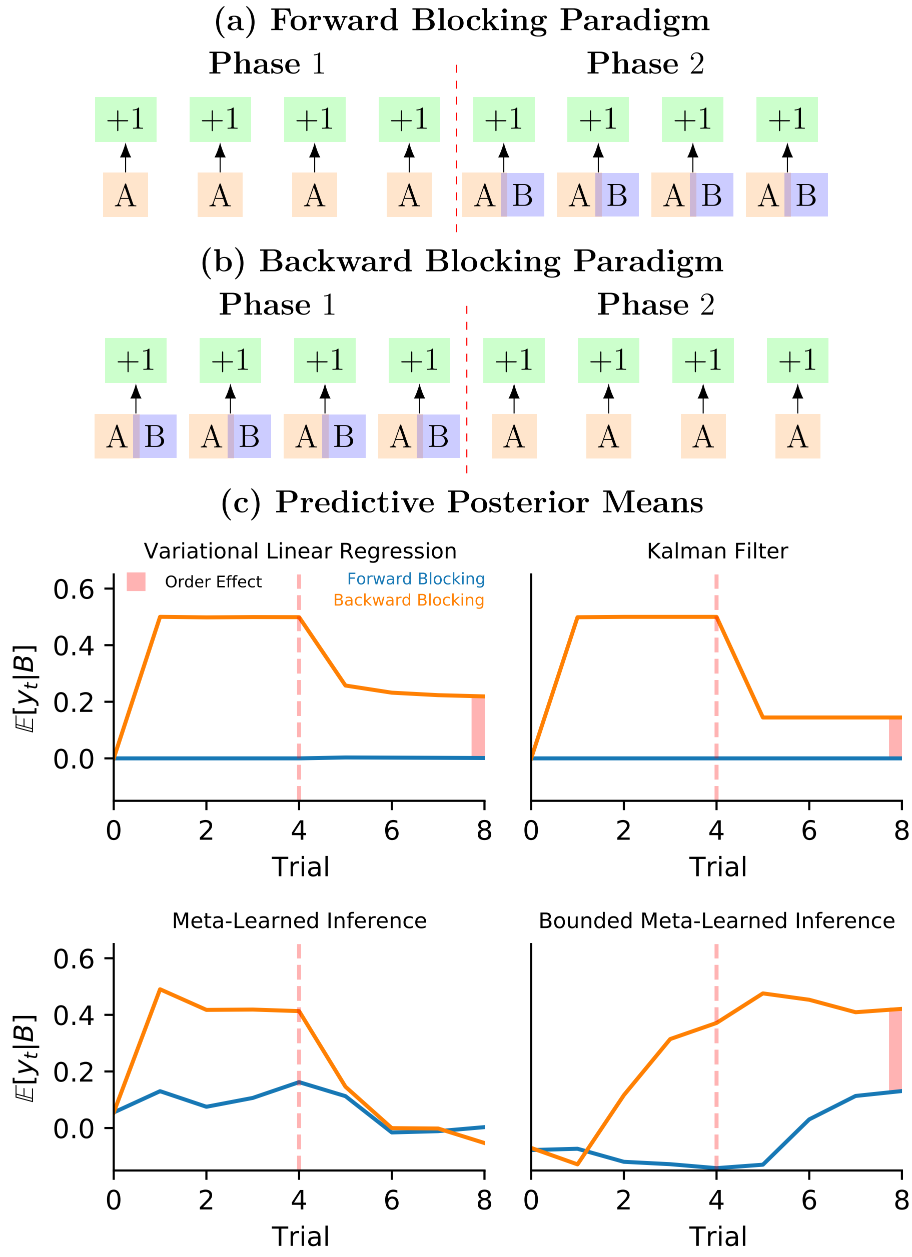


Figure 5.1: (a) Graphical illustration of the forward blocking paradigm. (b) Graphical illustration of the backward blocking paradigm. (c) Predictive posterior means for stimulus B during forward and backward blocking in different models. Kalman filters and variational linear regression assume an output variance of $\sigma = 0.1$, the Kalman filter additionally assumes a diffusion variance of $\tau = 1.0$. BMI uses a β-value of 0.001. All models except MI, which approximately simulates a stationary Kalman filter, show the empirically observed order effect.

Highlighting

In highlighting the compound stimulus AB is associated with outcome R, while the compound stimulus AC is associated with outcome S. Both associations are presented an equal amount of times. However, $AB \rightarrow R$ is presented predominantly in the first phase, while $AC \rightarrow S$ is presented predominantly in the second phase. Here, we follow the variant of (Daw et al., 2008) and set R to 1 and S to -1 and again use four observations per phase. This highlighting paradigm is illustrated graphically in Figure 5.2 (a).

When tested on A at the end of the second phase people tend to predict an outcome that is closer to 1. This is a primacy effect because A was predominantly rewarded early on. However, when tested on BC at the end of the second phase people tend to predict an outcome that is closer to -1. This is a recency effect because the association of C with -1 was more prevalent in the second phase. Because A has been paired with positive and negative rewards an equal number of times overall, any deviations from the null response after the second phase have to be attributed to the specific ordering of trials; the same holds for both parts of the compound BC.

Figure 5.2 (b) demonstrates that only variational linear regression and BMI are in agreement with the empirical data from the literature. Both are more likely to predict that A results in a positive outcome and that BC results in a negative outcome at the end of the second phase. Neither MI nor the Kalman filter can account for the empirical observations made in the highlighting paradigm. As more data is observed, both of them conclude that A is entirely unpredictive for the obtained reward and thus attribute all positive rewards to B and all neg-

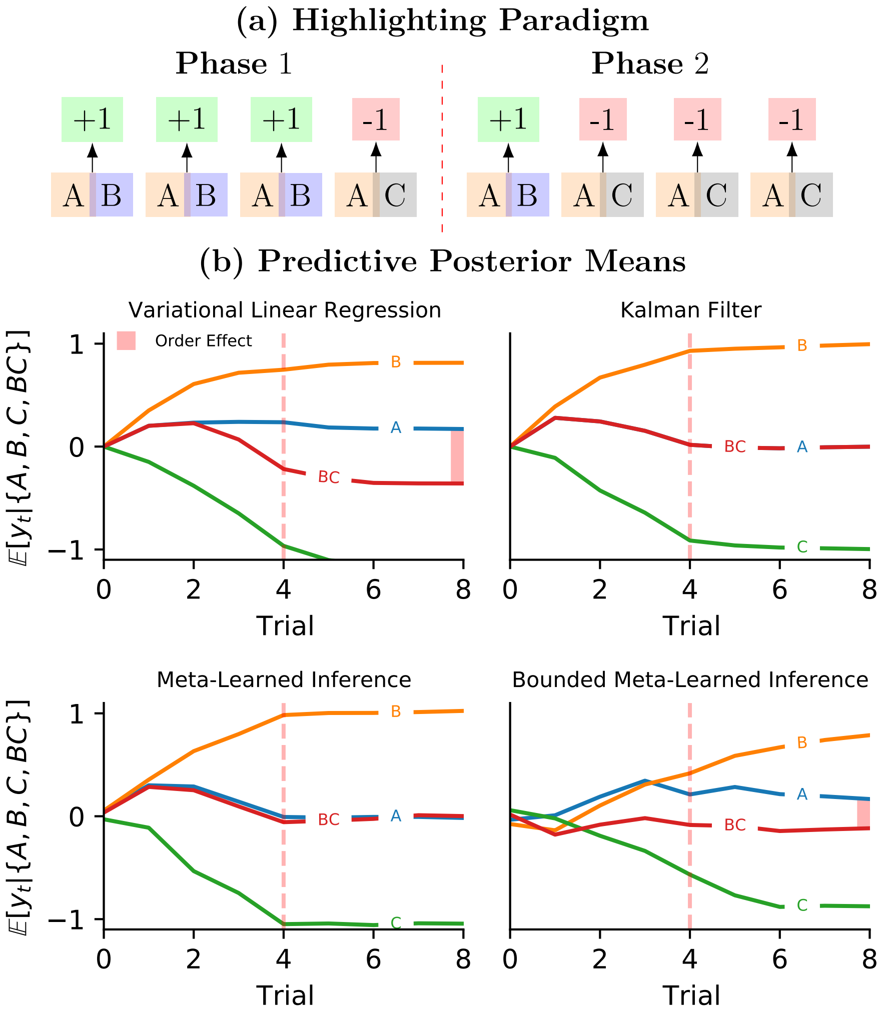


Figure 5.2: (a) Graphical illustration of the highlighting paradigm. (b) Predictive posterior means for different stimuli during learning in different models. Each line is labelled with the corresponding stimulus. Kalman filters and variational linear regression assume an output variance of $\sigma = 0.1$, the Kalman filter additionally assumes a diffusion variance of $\tau = 1.0$. BMI uses a β-value of 0.001. Both variational linear regression and BMI show the empirically observed order effect, whereas the Kalman filter and MI do not.

ative rewards to C. In turn, this also implies a null response to BC.

These results show that the hypothesis that people assume a non-stationary environment does not offer an explanation for order effects observed in the highlight paradigm. Meanwhile, both other explanations – inferences within a simplified space of hypotheses and making optimal use of limited computational resources – are able to account for effects in both forward/backward blocking and highlighting.

5.3.2 Function Learning

Next, we are going to examine how well each of the discussed theories generalizes beyond the setting of associative learning. From the function learning literature, we know that sequences with structured inputs are easier to learn than sequences with random order (Byun, 1996). Byun (1996), for example, contrasted how people learn different functions with one-dimensional inputs that are either presented randomly or in ascending order. Looking at linear, quadratic, and sinusoidal functions, they found that people tend to learn these functions better if inputs are presented in ascending order.

We set up an analogous learning environment: models observe ten inputs with their corresponding targets, and subsequently make predictions about ten additional inputs without feedback. To-be-learned relationships are based on either linear, quadratic or sinusoidal functions and tasks are generated as follows:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \prod_{t=1}^{T} \mathcal{N}(y_t|\mu_t, 0.1) \tag{5.13}$$

$$\mu_t = \begin{cases} \mathbf{w}_1 + \mathbf{w}_2 \cdot x_t & \text{if } c = 0 \\ \mathbf{w}_1 + \mathbf{w}_2 \cdot x_t + \mathbf{w}_3 \cdot x_t^2 & \text{if } c = 1 \\ \mathbf{w}_1 + \cdot \sin\left(2\pi \mathbf{w}_2 \cdot x_t + \mathbf{w}_3\right) & \text{if } c = 2 \end{cases} \tag{5.14}$$

$$p(x_{1:T}) = \prod_{t=1}^{T} \mathcal{N}(x_t|0, 1) \tag{5.15}$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) \tag{5.16}$$

where c is a uniformly distributed categorical variable that is sampled at the beginning of each task and kept constant for its entire duration. Input sequences of the initial phase being either sorted in ascending order or left unchanged. Examples for possible input sequences and all function types are illustrated in Figure 5.3 (a).

The same distribution over tasks is used to adapt our meta-learning models.[*] As in the associative learning simulations, this training distribution implies that MI should be invariant to rearrangements of data-points and thus that it should not differ in its generalization performance between the random and the sorted condition.

The primary quantity we are interested in is generalization performance during the second phase. Both variational linear regression and Kalman filters are

[*] Inputs that correspond to the targets from the previous step are masked in the second phase (set to an uninformative value of zero). This ensures that no feedback is provided to the meta-learning models after the end of the first phase.

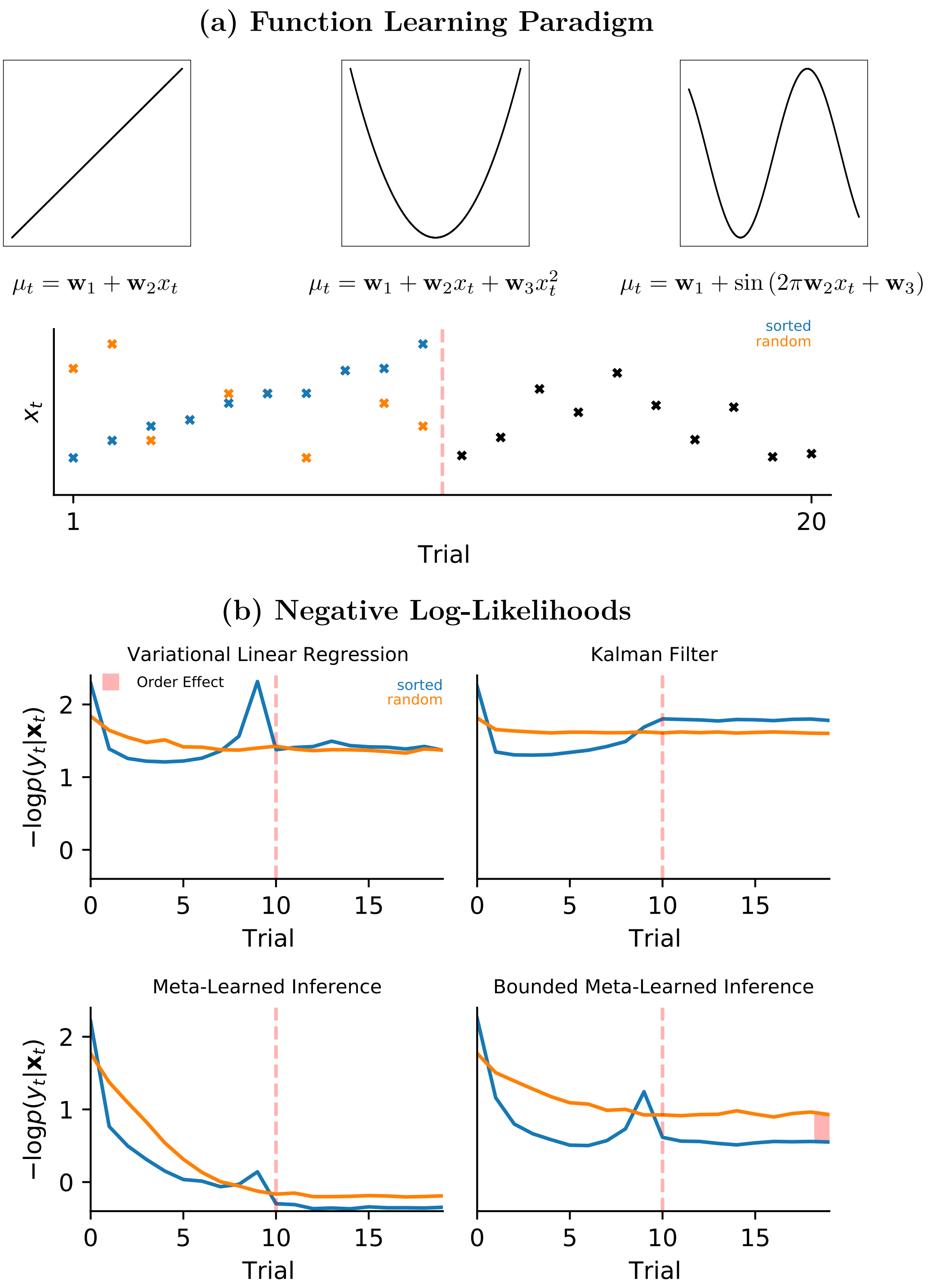


Figure 5.3: (a) Graphical illustration of the function learning paradigm. The upper panel shows examples of the three functions types used in our model simulations. The lower panels shows examples of input sequences for both the random and the sorted condition. (b) Negative log-likelihoods for different models plotted over number of observed data-points. Lower values correspond to better performance. Kalman filters and variational linear regression assume an output variance of $\sigma = 1.0$, the Kalman filters additionally assumes a diffusion variance of $\tau = 1.0$. BMI uses a β-value of 0.001. BMI is the only approach showing the empirically observed order effect. Kalman filters show the reverse effect, whereas the other models show no effect.

linear models, and thus the non-linear functions used in this task pose a challenge for them. Figure 5.3 (b) indicates that the performance of variational linear regression is identical for both conditions. Interestingly, the Kalman filter shows the reverse of the empirically observed order effect: generalization performance is worse with ascending order than it is with random order. Fitting a linear function to the most recent inputs – as done by Kalman filter – will worsen the fit to previous inputs. As inputs in the middle of sorted sequences are most likely according to the data generating distribution, neglecting these inputs in favor of more recent ones leads to reduced generalization performance in the second phase.

In MI, we find a minuscule preference for sorted sequences. Although MI should implement an algorithm that is invariant to trial order, this argument is still subject to a sufficiently complex model architecture. This finding indicates that MI requires a slightly more expressive model architecture to perform optimally in the given task. From all the models under consideration, only BMI fully captures the empirically observed order effect by showing a substantial improvement on sequences with ascending order. This is the case because a simple learning algorithm may still learn well if there exist structure in the data, whereas learning becomes more difficult if structure is absent. The simplifications made in variational linear regression, however, go too far; a linear relationship is not expressive enough for the given task.

Previously, we have seen that order effects in associative learning can be explained by different theories. However, moving to a non-linear function learning setting provides a challenge for some of them, including variational linear regression and Kalman filters. In contrast to this, BMI adapts flexibly to the non-linear function learning environment and also accounts for order effects found in function learning studies.

5.3.3 Multi-Task Learning

Not only are people sensitive to the presentation order of inputs, but they are also sensitive to the presentation order of different tasks (Lee et al., 1992, Flesch et al., 2018). For instance, when learning about multiple tasks simultaneously people tend to perform better when tasks are grouped together compared to an interleaved presentation order. Flesch et al. (2018) demonstrated this effect in an experimental study where people learned to categorize naturalistic images of trees according to one of two orthogonal task rules.

To investigate this effect, we created a simplified version of the experiment used by Flesch et al. (2018). Each episode starts with an initial phase of ten trials from two tasks, with feedback provided after each prediction. This initial phase is followed by a second phase with another ten trials from each task without feedback. Instead of using naturalistic images as done by Flesch et al. (2018), we directly provide a two-dimensional feature representation as inputs to our models. Specifically, tasks are generated using the following expression:

$$p(y_{1:T}|\mathbf{x}_{1:T}, \mathbf{w}) = \prod_{t=1}^{T} \mathcal{N}(y_t|\mu_t, 0.1) \tag{5.17}$$

$$\mu_t = \begin{cases} \mathbf{w}_A^T \mathbf{x}_t & \text{if task id} = A \\ \mathbf{w}_B^T \mathbf{x}_t & \text{if task id} = B \end{cases} \tag{5.18}$$

$$p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{x}_t|\mathbf{0}, \mathbf{I}) \tag{5.19}$$

$$p(\mathbf{w}_A) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) \tag{5.20}$$

$$p(\mathbf{w}_B) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{I}) \tag{5.21}$$

In the initial phase, tasks are either presented interleaved with each other or in blocks (i.e. ten presentations of task A, followed by ten presentations of task B). In the second phase, tasks are always presented interleaved with each other. The task identity can be observed by the learner. MI and BMI do so by receiving an additional one-hot vector corresponding to the current task identity. For all tasks we train separate models for the different conditions. This multi-task learning paradigm is illustrated in Figure 5.4 (a).

All meta-learning models are trained on the same distribution over tasks.[†] Again, this implies that MI should be invariant to rearrangements of data-points and thus that it should not differ in its generalization performance between the blocked and the interleaved condition.

The primary measurement of interest is again generalization performance on

[†] Inputs that correspond to the targets from the previous step are masked in the second phase (set to an uninformative value of zero). This ensures that no feedback is provided to the meta-learning models after the end of the first phase.

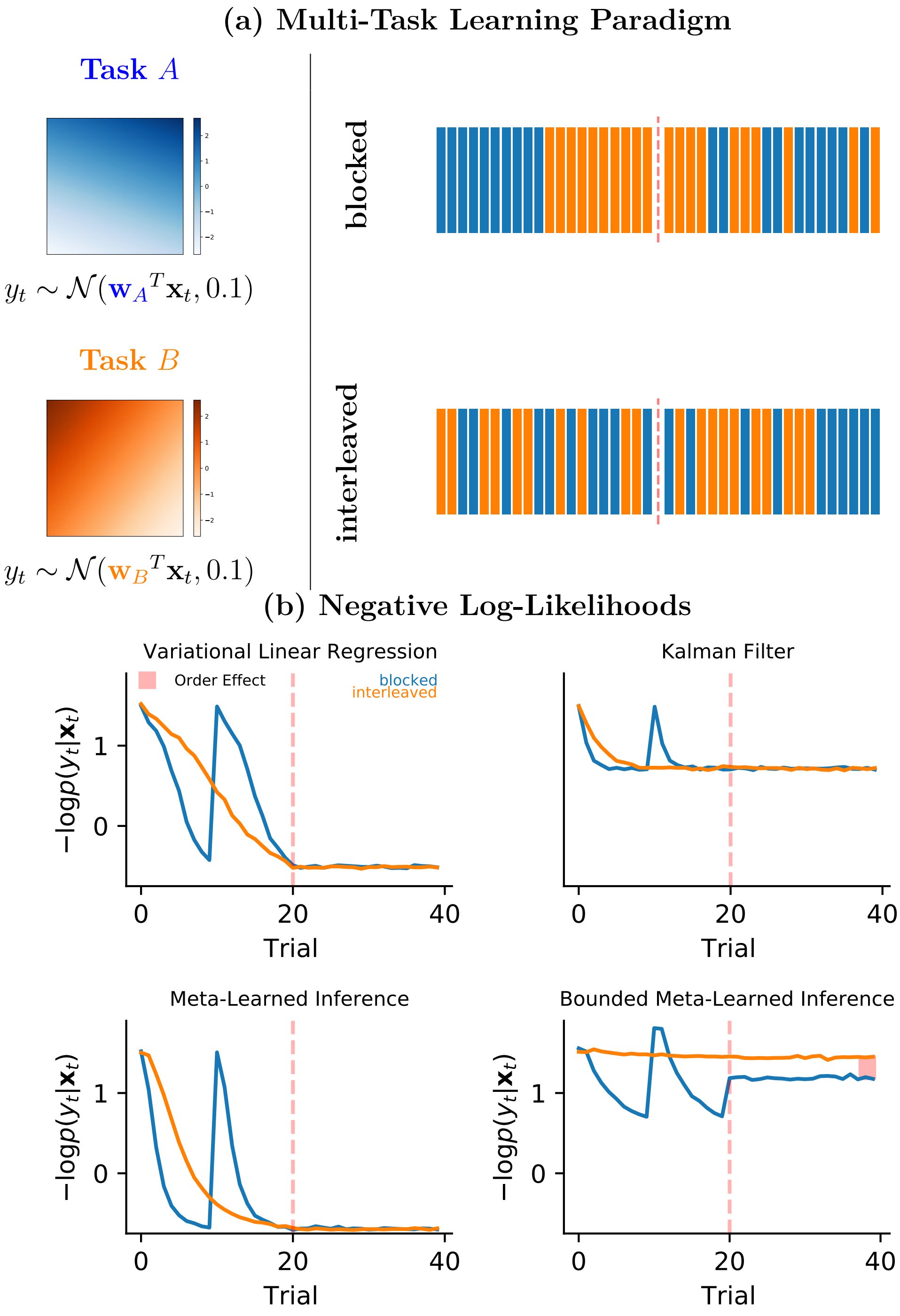


Figure 5.4: (a) Graphical illustration of the blocked and interleaved condition in the multi-task learning paradigm. Bars indicate the corresponding task (blue for task A, orange for task B). (b) Negative log-likelihoods for different models plotted over number of observed data-points. Lower values correspond to better performance. Kalman filters and variational linear regression assume an output variance of $\sigma = 0.1$, the Kalman filter additionally assumes a diffusion variance of $\tau = 1.0$. BMI uses a β-value of 0.001. BMI is the only approach showing the empirically observed order effect, i.e. its performance is better after a blocked presentation of tasks.

data-points in the second phase. There are multiple conceivable ways to extend single-task models (variational linear regression and Kalman filters) to the multi-task setting. The most natural choice, which we also adopt here, is to keep a separate model for each task and assume the ability to switch to the correct model based on the provided task identity. With this approach, both variational linear regression and the Kalman filter observe the same data-points in both conditions, and hence they predict no difference in generalization performance between interleaved and blocked sequences. Figure 5.4 (b) verifies this hypothesis. MI also shows no difference between the two conditions because for an ideal observer – which MI simulates – it is irrelevant in which order trials are observed.

BMI is the only model that shows an improved generalization performance when tasks are grouped together – a result that echoes the empirical findings reported by Flesch et al. (2018). We hypothesize that presenting tasks in blocks avoids interference between them, which in turn requires a less complex learning algorithm to make successful inferences. There are a few additional interesting patterns in the learning curves of BMI. First, after the switch between tasks in the blocked condition, prediction error rises above predicting at chance level. This might be caused by an urge towards sticking to predictions acquired in the old task. Second, we observe that performance in the testing phase is worse compared to the level reached during training. This indicates an interference of the learned mappings when both tasks are interleaved.

5.4 Conclusion

We have analyzed several order effects that have been identified in previous experimental studies. Historically, these effects have been investigated independently within different sub-disciplines, each attempting to find their own explanations. In this chapter, we brought these separated research areas together and investigated whether different theories provide explanations for order effects that are found across domains.

Through multiple model simulations, we have demonstrated that only BMI offers a domain-general explanation for the order effects under consideration, whereas other previously suggested theories do not readily account for all of them. Besides its empirical support, BMI also offers several practical advantages. It can explain order effects that arise due to reasoning correctly in a particular environment *and* for those that are due to limited computational resources. While it may be possible to construct models based on variational inference or Kalman filters that can account for the investigated effects, doing so would require non-trivial adjustments. In contrast to this, BMI extends readily to new conditions without requiring any modifications to the model, which we argue is one of its biggest advantages.

Although BMI offers two potential explanations for order effects, limited computational resources alone were sufficient to explain all of the effects investigated in this chapter. In future work, it might be interesting to look at order effects that are neither explained through limited computational resources nor the jittering mechanism of Kalman filters, but that require more complex assumptions about the environment. It might also be interesting to test whether the other predictions made by BMI can be confirmed empirically. For example, as shown in the function learning simulations, we expect to find a decrease in performance for inputs that are unlikely according to the data generating distribution. Furthermore, as shown in the multi-task learning simulations, we expect to find that performance becomes worse than chance at transition points between blocks of tasks. Finding such patterns in empirical studies would provide further evidence in favor of BMI.

More generally, we believe that cross-discipline studies – like the one presented here – are important in the search for general principles of human cognition because they provide additional validity for computational theories. A general theory of human learning does not only have to capture a specific aspect of human learning but should instead account for a wide range of different phenomena. In the context of this thesis, we demonstrate that BMI also explains why people use decision-making strategies (ref. Chapter 4) and that variants build for the reinforcement learning setting account for individual differences in human exploration strategies (ref. Chapter 6). These results add credibility to the idea that meta-learning and resource-rationality are domain-general principles of human learning.

Finally, our work also has implications for research in machine learning and artificial intelligence. Real-world environments are structured in many different ways and people are, in contrast to current machine learning models, very efficient at exploiting these structures for their benefits. If our goal is to build more human-like machines, it seems necessary to consider structured training environments that induce learning curricula (Bengio et al., 2009) along with models that can exploit these. Inevitably, such models have to be sensitive to the arrangement of data-points, ideally in a way that is similar to what people show. We have demonstrated that BMI is such a model and hence it could serve as the foundation for future research in this direction.

6

Where Do Exploration Strategies Come From?

Abstract: People constantly face the decision of whether they should exploit their currently available knowledge or whether they should instead explore parts of their environment they do not know much about yet. The reinforcement learning framework offers many approaches that address this trade-off between exploration and exploitation. Looking at human exploration on a two-armed bandit problem, we find that people apply several different exploration strategies. This leads us to the question of why people use these particular strategies. We hypothesize that people explore by making optimal use of limited computational resources and test this conjecture with the help of RL^3. We find that RL^3 displays characteristics that resemble individual differences between human participants and that it explains empirical data better than any other strategy under consideration.

This chapter is based on the following publication:

Binz, M., & Endres, D. (2019). Where Do Heuristics Come From?. In Proceedings of the 41th Annual Conference of the Cognitive Science Society (pp. 1402-1408).

The code to reproduce the presented results is publicly available under https://github.com/marcelbinz/MDLDQN.

6.1 Introduction

Knowing how to efficiently balance between exploring unfamiliar parts of an environment and exploiting currently available knowledge is a requirement for any intelligent organism. *Multi-armed bandits* offer an idealized problem setting that allow us to study the trade-off between exploration and exploitation. In a multi-armed bandit problem, an agent repeatedly interacts with k slot machines, each providing noisy rewards according to some unknown probability distribution (Lattimore and Szepesvári, 2020).

To behave optimally in such problems, an agent has to maximize the total amount of accumulated rewards. It is possible to express Bayes-optimal strategies as the result of a planning process in an augmented MDP (Duff and Barto, 2002). However, analytical solutions are only available for a few special cases; for example, when considering infinite time horizons and geometric discounting, which results in the *Gittins index strategy* (Gittins, 1979). The general intractability of the problem led to the development of countless heuristic approaches for solving the exploration-exploitation trade-off; including approaches based on sampling (Russo et al., 2017), visitation frequencies (Auer et al., 2002), uncertainty bonuses (Kaufmann et al., 2012), and information gain (Russo and Van Roy, 2014).

The sheer magnitude of available exploration strategies begs the question: how do people explore? Prior work indicated that we explore intelligently by using uncertainty estimates to guide our choices (Speekenbrink and Konstantinidis, 2015, Wu et al., 2018, Gershman, 2019, Schulz and Gershman, 2019, Schulz et al., 2019), but also that our choices systematically deviate from the ones prescribed by the Bayes-optimal strategy (Steyvers et al., 2009, Zhang and Angela, 2013).

Is there any justification for why people use particular exploration strategies? We approach this question from the perspective of resource-rationality (Simon, 1990a, Gershman et al., 2015, Lieder and Griffiths, 2020), and hypothesize that people attempt to explore optimally but are subject to limited computational resources. To test this conjecture, we make use of RL^3. RL^3 is a meta-learned approximation to the Bayes-optimal strategy that accounts for limited computational resources. RL^3 parametrizes the to-be-learned algorithm by an RNN, which is trained via meta-learning (Duan et al., 2016, Wang et al., 2016) to implement a reinforcement learning algorithm that (1) explores optimally, and (2) is as simple as possible. Modifying the relative importance of the two factors leads to a spectrum of resource-rational algorithms, each possessing different properties. Algorithms without resource limitations approximate the Bayes-optimal strategy, whereas more constrained algorithms must implement simpler exploration strategies.

We find that RL^3 displays characteristics of human exploration on both a qualitative and quantitative level. It not only captures individual differences in human exploration strategies but also explains empirical data better than any other strategy under consideration. Taken together, these results indicate that the seemingly sub-optimal exploration strategies used by people might be a consequence of the constraints under which these very strategies are learned.

6.2 Computational Models

Multi-armed bandits are MDPs consisting of a single state. In each time-step t, an agent selects one out of k available actions and is rewarded according to an unknown distribution based on its choice. The agent's objective is to maximize its total sum of rewards during T interactions with the problem. Finding the policy that optimally balances exploration and exploitation in such problems is incredibly difficult. In the special case of an infinite horizon and geometric discounting, the Bayes-optimal solution is the Gittins index strategy (Gittins, 1979). More generally, the Bayes-optimal solution is defined as the result of a planning process in an augmented MDP (Duff and Barto, 2002).

The difficulty of the multi-armed bandit problem led to the development of several heuristic approaches for addressing the exploration-exploitation trade-off. These methods can roughly be categorized into two major groups: *directed* and *random* exploration strategies. Directed exploration attempts to gather information about uncertain but learnable parts of the environment, while random exploration injects stochasticity of some form into the policy. Gershman (2018) showed that these two principles can be distinguished exactly under certain conditions. Next, we present several algorithms for interacting with bandit problems. We first summarize the random and directed exploration strategies suggested by Gershman (2018). Then, we describe how RL^3 can be applied to the given problem setting.

6.2.1 Value-Directed Exploration

Following Gershman (2018), we consider two-armed bandit problems with normal distributions over the mean of rewards for each arm θ_a, and a normal distribution over the reward at each time-step r_t:

$$p(\theta_a) = \mathcal{N}(\theta_a | \mu_{0,a}, \sigma_{0,a}) \tag{6.1}$$

$$p(r_t | a_t) = \mathcal{N}(r_t | \theta_{a_t}, \tau) \tag{6.2}$$

If our objective is to maximize the total sum of rewards, keeping track of how rewarding each arm is, is a good starting point. In our case, agents maintain a posterior distribution over mean rewards for each arm. Because everything is normally distributed, this posterior will also be normally distributed, i.e. $p(\theta_{t+1}|r_t, a_t) = \mathcal{N}(\theta_{t+1}|\mu_{t+1}, \sigma_{t+1})$. The corresponding updating equations are given by:

$$(\mu_{t+1,a}, \sigma^2_{t+1,a}) \leftarrow \begin{cases} (\mu_{t,a}, \sigma^2_{t,a}), & \text{if } A_t \neq a \\ \left(\mu_{t,a} + \alpha\left(r_t - \mu_{t,a}\right), \sigma^2_{t,a} - \alpha\sigma^2_{t,a}\right), & \text{if } A_t = a \end{cases} \tag{6.3}$$

$$\alpha \leftarrow \frac{\sigma^2_{t,a}}{\sigma^2_{t,a} + \tau^2_a} \tag{6.4}$$

Let us now the define the following quantities:

$$\begin{aligned} V_t &= \mu_{t,0} - \mu_{t,1} \\ RU_t &= \sigma_{t,0} - \sigma_{t,1} \\ TU_t &= \sqrt{\sigma^2_{t,0} + \sigma^2_{t,1}} \end{aligned} \tag{6.5}$$

V_t constitutes the estimated difference in value between both arms, while RU_t and TU_t describe relative and total uncertainty. These quantities allow us to formulate many popular exploration strategies. Perhaps the simplest strategy is to select the arm with the higher expected reward. In other words: select arm 0 if $V_t > 0$ and arm 1 otherwise. This is a purely exploitation-based approach that never explores. A modified version of this idea is to choose the arm with the higher expected reward probabilistically. Gershman (2018) suggested to do so by transforming the value difference through the cumulative distribution function of a standard normal distribution $\mathbf{\Phi}$:

$$p(A_t = 0 | m = \text{VD}) = \mathbf{\Phi}(w_1 V_t) \tag{6.6}$$

where w_1 is a parameter that controls how noisy choices are. Equation 6.6 implements a type of random exploration, which we refer as *value-directed exploration.*

6.2.2 Upper Confidence Bounds

Value-directed exploration is based on expected rewards alone. It turns out that we can explore more efficiently than that by using how uncertain the agent is. Intuitively, an agent can still learn a lot about arms with high uncertainty, whereas it is almost sure how much rewards arms with low uncertainty provide. *Upper confidence bound algorithms* (UCB, Auer et al., 2002) formalize this idea by adding an uncertainty-based bonus reward. This can be expressed using the previously defined quantities; in particular, the relative uncertainty RU_t:

$$p(A_t = 0|m = \text{UCB}) = \mathbf{\Phi}\left(w_1 V_t + w_2 RU_t\right) \tag{6.7}$$

If the uncertainty for arm 0 is higher than for arm 1, RU_t will be positive, which in turn leads to a higher probability of selecting arm 0. If uncertainty for arm 1 is higher, the agent will be encouraged to select arm 1. Equation 6.7 is thus a directed exploration strategy – it *directs* an agents towards arms with high uncertainty. The parameters w_1 and w_2 determine the influence of value difference and relative uncertainty, respectively.

6.2.3 Thompson Sampling

Thompson sampling (Russo et al., 2017) is another way to incorporate uncertainty estimates into the decision-making process. An agent that applies Thompson sampling selects arms relative to their probability of being optimal. This probability can be expressed analytically for bandits with two arms and normally distributed posteriors (Gershman, 2018):

$$\begin{aligned} p(A_t = 0|m = \text{TS}) &= p(\theta_{t,0} > \theta_{t,1}) && (6.8) \\ &= p(\theta_{t,0} - \theta_{t,1} > 0) && (6.9) \\ &= \mathbf{\Phi}\left(\frac{V_t}{TU_t}\right) && (6.10) \end{aligned}$$

Like value-directed exploration, Thompson sampling is a random exploration strategy. It also probabilistically selects the arm with the higher expected reward. However, in contrast to value-directed exploration, it scales the value difference by the total uncertainty TU_t. This implies that behavior becomes more stochastic if uncertainties are high in general.

6.2.4 Hybrid Model

It is possible to combine all three algorithms presented so far in a single, unified probit regression model:

$$p(A_t = 0 | m = \text{HYBRID}) = \mathbf{\Phi}\left(w_1 V_t + w_2 RU_t + w_3 \frac{V_t}{TU_t}\right) \tag{6.11}$$

where $\mathbf{w} = [w_1, w_2, w_3]$ define the influence of each exploration strategy. For $\mathbf{w} = [w_1, 0, 0]$ we recover value-directed exploration, for $\mathbf{w} = [w_1, w_2, 0]$ we recover UCB, and for $\mathbf{w} = [0, 0, 1]$ we recover Thompson sampling.

Fitting the coefficients of the hybrid model to empirical data allows us to inspect how much a given agent relied on value-directed, directed, and random exploration respectively. Figure 6.1 verifies that this procedure is able to recover different exploration strategies from data generated by them. We will apply this form of analysis to both human participants and to meta-learned agents in the next section.

6.2.5 RL^3

Finally, we want to look at how meta-learning can be used to discover different exploration algorithms for the given two-armed bandit problem. Bandits are MDPs with a single state, and thus we can directly apply the RL^3 method described in Section 3.2. RL^3 parametrizes the to-be-learned exploration algorithm with a RNN. The RNN takes previous actions and rewards from a ban-

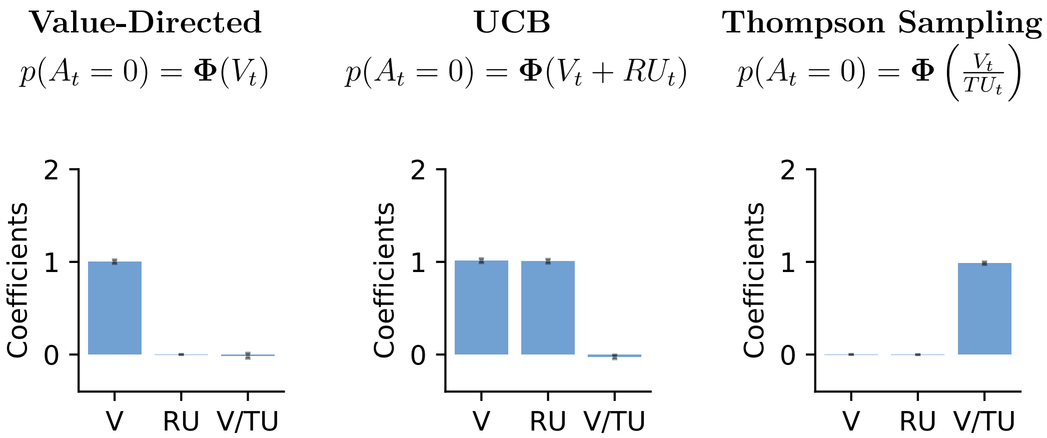


Figure 6.1: Parameters obtained by fitting the hybrid probit regression model to data generated by value-directed, directed (UCB), and random exploration (Thompson sampling). The figure highlights that this form of analysis is able to recover each exploration strategy from data generated by it.

dit problem as inputs, making the output a function of the entire history $h_t = a_{1:t-1}, r_{1:t-1}$. The outputs of the network parametrize a distribution over action-values of the optimal policy. A good algorithm has to integrate information from the history to accurately predict action-values of the optimal policy, based on which it subsequently selects the appropriate strategy.

Initially, the RNN implements a random mapping. During meta-learning it is then turned into a resource-rational reinforcement learning algorithm. This is accomplished by minimizing the RL^3 objective until convergence:

$$\mathcal{L}_{\text{RL}^3}(\mathbf{\Lambda}) = \mathbb{E}_{q(\mathbf{\Theta}|\mathbf{\Lambda})}\left[\mathbb{E}_{p(a_{1:T}, r_{1:T})}\left[\sum_{t=0}^{T-1} - \log p(q_{t+1}|h_{t+1}, a_{t+1}, \mathbf{\Theta})\right]\right] + \beta \text{KL}\left[q(\mathbf{\Theta}|\mathbf{\Lambda})||p(\mathbf{\Theta})\right] \tag{6.12}$$

Here, we use an approximation to the Q-Learning targets q_t defined in Equation 3.15. This approximation is computed using a single sample from the encoding distribution (Lipton et al., 2017) of a separate target network (Mnih et al., 2015). The target network is synchronized with the main network every 100 training iterations.

The RNN implements a resource-rational reinforcement learning algorithm through its recurrent activations after meta-learning is completed. The outputs of the network approximate the action-value function of the optimal policy, from which the corresponding policy can be derived. In Appendix C we provide a full specification on the network architecture, meta-learning procedure and choice of prior.

RL^3 may implement different exploration strategies depending on its β-value. Duan et al. (2016) demonstrated that a similar algorithm without resource limitations closely approximates the Bayes-optimal policy for bandit problems of low to medium complexity. Increasing the β-value will lead to algorithms with a shorter description length, which in turn implement simpler exploration strategies. In the following section, we investigate whether we can understand individual differences in human exploration by considering models with different β-values.

6.2.6 Model Summary

Let us briefly summarize the presented exploration strategies before showing how they can help us to understand how people explore:

Value-directed exploration Probabilistically selects the arm with the higher estimated reward. It only depends on the value difference and does not take into account how uncertain the agent is.

UCB Implements the idea of optimism in the face of uncertainty by adding a bonus reward that directs the agent towards arms with high uncertainty.

Thompson sampling Samples arms relative to their probability of being optimal. In contrast to value-directed exploration, it does utilize uncertainty estimates as a scaling factor that determines how stochastic choices are.

Hybrid model Integrates value-directed exploration, UCB and Thompson sampling into a single exploration strategy.

RL^3 A meta-learned algorithm that makes optimal use of limited computational resources. The implemented exploration strategy will depend on the relative weighting between performance and cost for computational resources.

6.3 Empirical Analysis

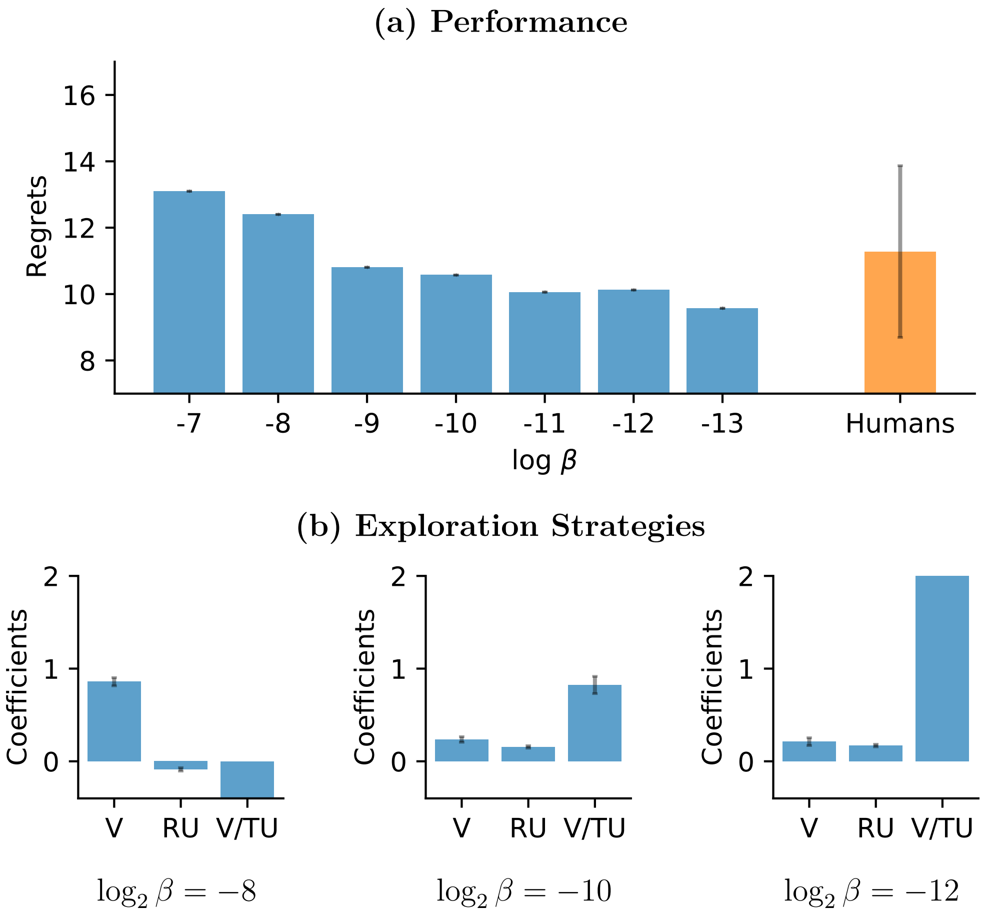
Next, we demonstrate how the previously described models can help us to understand which exploration strategies people are using. Our analysis involved three parts. First, we fitted the probit regression coefficients of the hybrid model (ref. Equation 6.11) to data generated by RL^3. This allowed us to reveal how much a given meta-learning agent relied on value-directed, directed, and random exploration. Then, we performed the same analysis for data generated by people, and compared the resulting coefficients to those obtained from RL^3. Finally, we ran a Bayesian model comparison to obtain a quantitative measure that describes how well the considered strategies capture human exploration.

For our analysis, we relied on data collected by Gershman (2018), which contains records of 44 participants, each playing 20 two-armed bandit problems with an episode length of $T = 10$. The mean reward for each arm was drawn from $\mathcal{N}(\theta_a|0, 10)$ at the beginning of an episode and the reward in each step from $\mathcal{N}(r_t|\theta_{a_t}, 1)$.

6.3.1 Model Simulations

First, we illustrate that RL^3 leads to the emergence of a spectrum of diverse exploration strategies. We trained otherwise identical models with varying regularization factors β on the two-armed bandit problem described above until convergence. Reported results are averaged over 5 random seeds unless otherwise noted. Figure 6.2 (a) shows that performance continuously improved as β decreased, which confirmed our expectation that RL^3 should become better at solving the exploration-exploitation trade-off when it faces fewer resource constraints.

Fitting the aforementioned probit regression model to data generated by RL^3 revealed value-based characteristics towards one end of the spectrum of resource-rational algorithms as shown in Figure 6.2 (b). Towards the other end of the spectrum, we observed a transition towards Thompson sampling-like strategies,


(a) Performance

(b) Exploration Strategies

$\log_2 \beta = -8$ $\log_2 \beta = -10$ $\log_2 \beta = -12$

Figure 6.2: Results for RL^3 with different β-values. (a) Visualization of per episode regret averaged over 5 models and 1000 episodes. Lower regret indicates a better performance. (b) Coefficients of the probit regression analysis for RL^3 with different β-values. Error bars indicate uncertainties (one standard deviation) in the coefficients estimated through a Laplace approximation.

with smaller influences of value-directed and directed characteristics.

6.3.2 Human Exploration Strategies

We performed the same probit regression analysis for each participant that took part in the study of Gershman (2018). To get a better understanding of the exploration strategies people used, we applied a dimensionality reduction technique to the resulting probit regression coefficients. The results are visualized in Figure 6.3 (a). This analysis revealed a continuum of strategies within the population. We performed an additional cluster analysis and visualized coefficients of three example participants in Figure 6.3 (c). While some participants seemed to adopt Thompson sampling (cluster 2), others relied on UCB (cluster 1) or a mixture between both (cluster 3).

We then compared the probit regression coefficients of participants to the ones of RL^3. Figure 6.3 (b) visualizes coefficients for 35 models (5 for each value of β) alongside those of human participants. Although some parts of the embedding were over- and underrepresented, the overall variability of human exploration strategies was captured well by the set of RL^3 models.

6.3.3 Model Comparison

To obtain a quantitative measure for the explanatory power of RL^3, we performed a Bayesian model comparison. Appendix D provides a detailed description of the methods we used for statistical analysis. In particular, we computed the posterior probability that a participant used a given strategy. Figure 6.4 shows posterior probabilities for each participant and model. We found that RL^3 was best at capturing participants' choices in 26 out of 44 participants. In 21

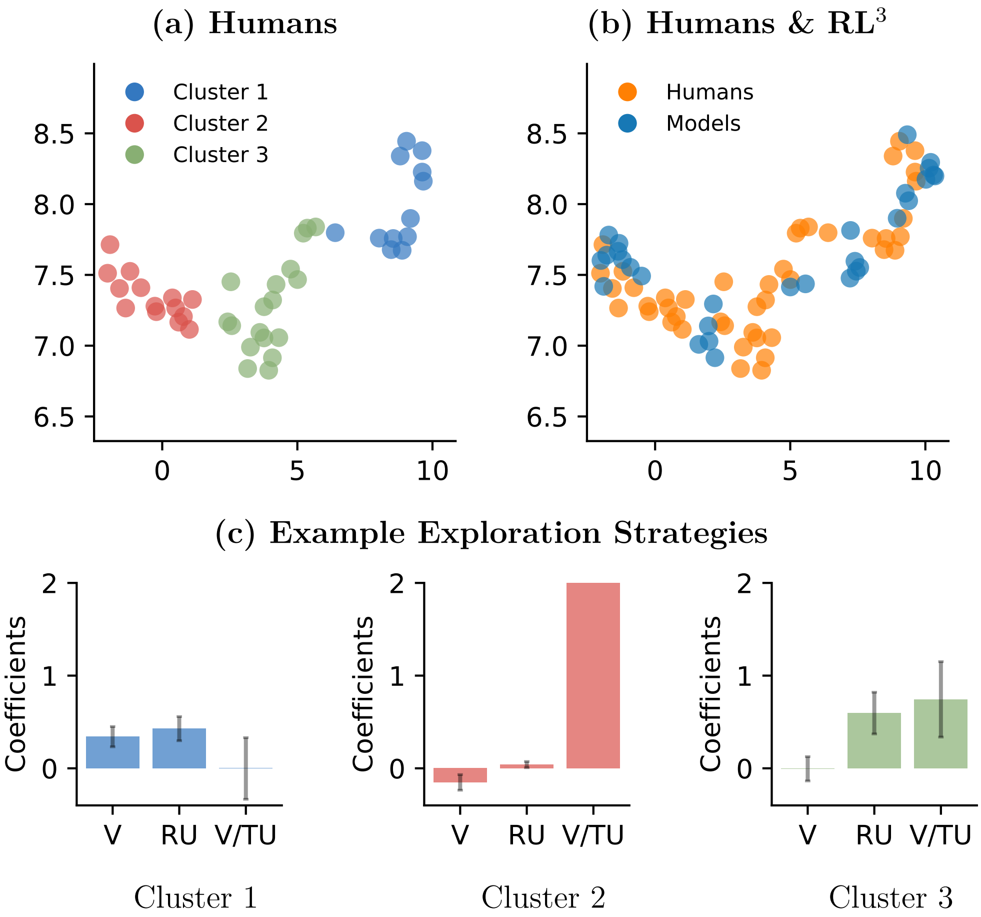


Figure 6.3: Visualization of human exploration strategies alongside those from RL^3. (a) UMAP (McInnes et al., 2018) embedding of probit regression coefficients for all participants. We also show the result of a cluster analysis obtained from a mean-shift clustering. (b) Joint UMAP embedding of coefficients for human participants and RL^3. (c) Probit regression coefficients of example participants from each cluster. Error bars indicate uncertainties (one standard deviation) in the coefficients estimated through a Laplace approximation.

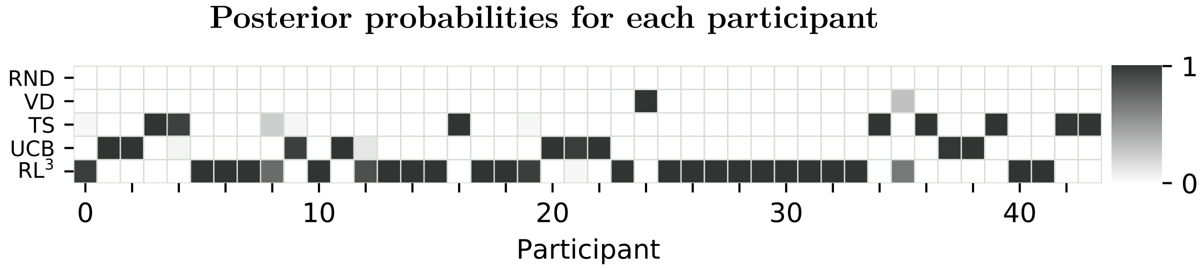


Figure 6.4: Posterior distributions for each participant over different exploration strategies. High values indicate that the participant was likely to use the corresponding strategy.

of those the model evidence decisively favored RL^3 ($p(m = \text{RL}^3|\mathcal{D}_i) > 0.99$). Amongst the participants not best described by RL^3, nine were best described by directed exploration (UCB), eight by random exploration (TS), and one by value-directed exploration (VD). The model evidence on the aggregated data of all participants indicated that RL^3 was overall better at capturing human exploration strategies than any other strategy under consideration ($p(m = \text{RL}^3|\mathcal{D}) \approx 1$). We also observed that different participants were described best by different β-values. There were 24 participants best described by $\log_2(\beta) = -7$, ten by $\log_2(\beta) = -8$, three by $\log_2(\beta) = -9$, four by $\log_2(\beta) = -10$, two by $\log_2(\beta) = -11$, and one by $\log_2(\beta) = -12$.

6.4 Conclusion

We hypothesized that the idea of resource-rationality offers a justification for the seemingly sub-optimal exploration strategies used by people in multi-armed bandit problems. To test this hypothesis, we compared human data to a family of meta-learned reinforcement learning algorithms called RL^3. RL^3 discovered a spectrum of exploration strategies that resembled human exploration without being explicitly trained to do so. Further model comparisons demonstrated that

RL^3 also described human exploration well on a quantitative level.

There are, however, a number of research questions left to be explored. For example, we plan to investigate whether manipulating β-values in RL^3 results in behavior that aligns with what we find when resources are manipulated in people. In this context, it might be interesting to look at data from children (Schulz et al., 2019, Meder et al., 2020), people with brain dysfunctions (Bechara et al., 1994), and people who have been put under cognitive load (Cogliati Dezza et al., 2019).

The bandit setting does not contain any mechanism that enables an agent to control its environment. Thus, bandits only allow to study a limited aspect of human exploration; for instance, they certainly do not explain how young children explore their environment during play (Orhan et al., 2020). A full understanding of human exploration will require much richer paradigms. An important advantage of the presented method is that it can be applied to more complex domains without algorithmic modifications. Therefore, we view the multi-armed bandit problems studied in this chapter merely as the first step towards investigating exploration in more complex domains.

In the bigger context of this thesis, we have shown the ideas of meta-learning and resource-rationality do not only apply to supervised learning problems but also to reinforcement learning problems. Recent work on meta-learned reinforcement learning algorithms – similar to the one employed here – demonstrated that such systems develop capabilities that allow them to perform model-based planning (Wang et al., 2016), causal reasoning (Dasgupta et al., 2019), and few-shot learning (Santoro et al., 2016). Having scalable systems that are capable of such feats opens up new possibilities for the study of human cognition.

7

Discussion

The goal of this thesis was to establish general principles that drive human learning across domains. Identifying a minimal set of such principles is not only valuable for our understanding of how people learn but also for building more human-like machines. In particular, I have put forward three principles that I believe to be important. They are generalization, adaptation, and simplicity. There has already been a lot of prior work on them, so it is perhaps somewhat surprising that a domain-general model realizing all of them did not exist. To close this gap, I have presented a framework that combines meta-learning with the minimum description length principle.

In three different studies, I have shown that instantiations of this framework – BMI and RL^3 – captured many aspects of human learning across different domains. In the context of decision-making, BMI discovered previously suggested heuristics *and* selected between them appropriately. It also made precise predictions about if and when a particular heuristic should be used, which were subsequently confirmed in three new experiments. BMI also captured order effects across different domains, including associative learning, function learning, and multi-task learning, without requiring any modification. In the context of reinforcement learning, RL^3 discovered a spectrum of exploration strategies that resembled individual differences in human exploration, which demonstrated that presented ideas can also scale beyond supervised learning problems. In summary, the presented framework offers a domain-general, scalable, and empirically supported theory of human learning.

7.1 Limitations

It is also important to discuss what issues BMI and RL^3 do not address. First of all, I do not view them as process models; that is to say, they do not offer significant insights with regard to the processes that transform inputs into outputs. Instead, they focus on a question on the computational level of analysis: how would an optimal learner that is subject to limited computational resources behave in a particular environment? Having said that, I believe that pushing down these models towards the algorithmic and implementational level will allow for additional insights into the human mind. A potential path towards this goal is to make use of biologically plausible model architectures. Recently, there has been a lot of progress in training *spiking neural networks* (Maass, 1997, Bellec et al., 2020, Wunderlich and Pehle, 2020). If these algorithms can be successfully applied to the meta-learning setting, they would be the ideal candidate for this purpose.

BMI and RL^3 also only implement a specific notation of resource-rationality; one accounting for how many bits are required to store the learning algorithm. Presently, they do not account for memory or time constraints during the learning algorithm's execution, which are the objects studied in other resource-rational models (Ortega and Braun, 2013, Zaslavsky et al., 2018, Ho et al., 2020, Gershman, 2020, Sanborn et al., 2010, Vul et al., 2014, Lieder and Griffiths, 2017). Extending BMI and RL^3 to such types of resource-rationality is an interesting topic for future research. The natural way to do this involves placing information processing constraints on network activations in addition to the already existing constraints on its parameters. Whether the distinction between different types of resource constraints is relevant in practice remains to be seen.

Finally, all models presented in this thesis used a particular type of RNN to implement the learning algorithm. This was a heuristic modeling choice made out of convenience. For the sake of the resource-rational argument, the function approximator that optimally solves the trade-off between performance and description length should have been employed instead. Identifying this function approximator can – in principle – be part of the meta-learning objective by combining it with methods for automated *neural architecture search* (Elsken et al., 2018, Stanley et al., 2019). In this context, it would also be interesting to test whether architectures found by these methods contain sub-parts that can be mapped onto different brain areas.

7.2 Future Directions

The idea that people are adapted to a particular environment is a central premise of the meta-learning framework. The meta-learning models used in this thesis were adapted to tasks that could be encountered in the subsequent experiment. However, what we actually want to express is that people are adapted to the environment they live in, and not to the experiment some crazy psychologist come up with. Therefore we should ask ourselves: how can we construct meta-learning distributions that reflect real-world learning problems? I can see several possibilities to approach this question: (1) construct a meta-learning distribution based on a collection of real-world data-sets, (2) use crowdsourcing services to ask people to generate learning problems, or (3) generate learning problems automatically based on some objective. Each of these approaches has its own advantages and disadvantages. Future work should investigate if any of them can enrich our understanding of human learning.

People observe a rich stream of sensory information. Most computational models of human behavior – including those used in this thesis – abstract away much of this information and rely on idealized stimuli instead. In part, this is out of necessity; traditional models do not scale easily to naturalistic environments. However, the study of naturalistic environments is important because the "conclusions that are reached when experimenting with pared-down or idealised stimuli may be different from those reached when considering more complex or naturalistic data, since the simplicity of the stimuli can stifle potentially important emergent phenomena" (Hill et al., 2020). In principle, meta-learning is not subject to any fundamental scaling limitations. Indeed, it has already been applied to construct learning algorithms that process raw visual inputs (Santoro et al., 2016, Finn et al., 2017, Mishra et al., 2017, Gordon et al., 2018, Zintgraf et al., 2019a). Therefore, one of the primary goals of future work should be to apply meta-learning to study human behavior in more naturalistic environments.

Finally, there are also practical applications for the models presented in this thesis. For example, in computational psychiatry (Huys et al., 2016), they could help us identify which environmental characteristics cause particular psychiatric symptoms. The acquired insights could then subsequently be used to ask how the environment needs to be changed to treat these symptoms. The framework of self-play (Bansal et al., 2017, Silver et al., 2018) offers another possible application. The general idea of self-play is to set up multiple instances of an agent, which are then improved iteratively by pitting them against each other. While self-play is a powerful tool for obtaining agents that can cooperate or compete with each other, it often results in behavior that is unnatural from a human perspective (Carroll et al., 2019). If we want self-play agents to exhibit more human-like behavior, they would need to be exposed to human-like agents during self-play. The models presented in this thesis could be the first step in this direction.

7.3 Conclusion

What are the general principles that drive how people learn? In this thesis, I have set out to answer this question. Towards this end, I have put forward three principles that I believe to be important; generalization, adaptation, and simplicity. Together, these three principles revealed a lot of structure in what initially looked like sub-optimal human behavior. However, whether they also form a minimal set of principles – or if additional principles are needed – can only be answered by looking at a more extensive set of domains.

I began this thesis with two motivating examples: the hobby gardener who tries to figure out how to grow vegetables in his garden and someone who attempts to master the game of chess. The hobby gardener is clearly the better metaphor for the studies presented in this thesis, and arguably also for studies in cognitive science more generally. In part, this is because traditional computational models are confined to such a setting. I hope that the ideas presented here allow us to scale up computational models of human learning, such that one day they also can be applied to understand how the chess player improves his game.

A
Power Analysis

Environments with continuous features can facilitate statistical analysis as fewer trials are needed to observe expected effects. To verify this hypothesis, we conducted a power analysis for an environment with continuous features and one for an environment, where features are dichotomized based on their median. Here, we present results from environments with known feature rankings and $T = 10$ decisions per task.

In both settings, we computed how many tasks are on average required to distinguish the single cue heuristic from the ideal observer model, assuming that decisions are made by the single cue heuristic. In dichotomized environments, ties between features of two options are likely, and hence we modified the single cue heuristic to make decisions based on the first feature that discriminates between both options.

We assumed that decisions are made by the single cue heuristic and measured the average support for the single cue heuristic over the ideal observer model on a single task by computing log-Bayes Factors (Kass and Raftery, 1995) between both strategies:

$$\begin{aligned}\log B &= \mathbb{E}_{p(\mathbf{x}_{1:T}, y_{1:T})}\left[\sum_{t=1}^{T}\left[\int p(c_t|\mathbf{x}_t, w, m = \text{SC}) \log\left(\frac{p(c_t|\mathbf{x}_t, w, m = \text{SC})}{p(c_t|\mathbf{x}_t, \mathbf{w}, m = \text{IO})}\right) dc_t\right]\right] \\ &= \mathbb{E}_{p(\mathbf{x}_{1:T}, y_{1:T})}\left[\sum_{t=1}^{T} \text{KL}\left[p(c_t|\mathbf{x}_t, w, m = \text{SC})||p(c_t|\mathbf{x}_t, \mathbf{w}, m = \text{IO})\right]\right] \qquad \text{(A.1)}\end{aligned}$$

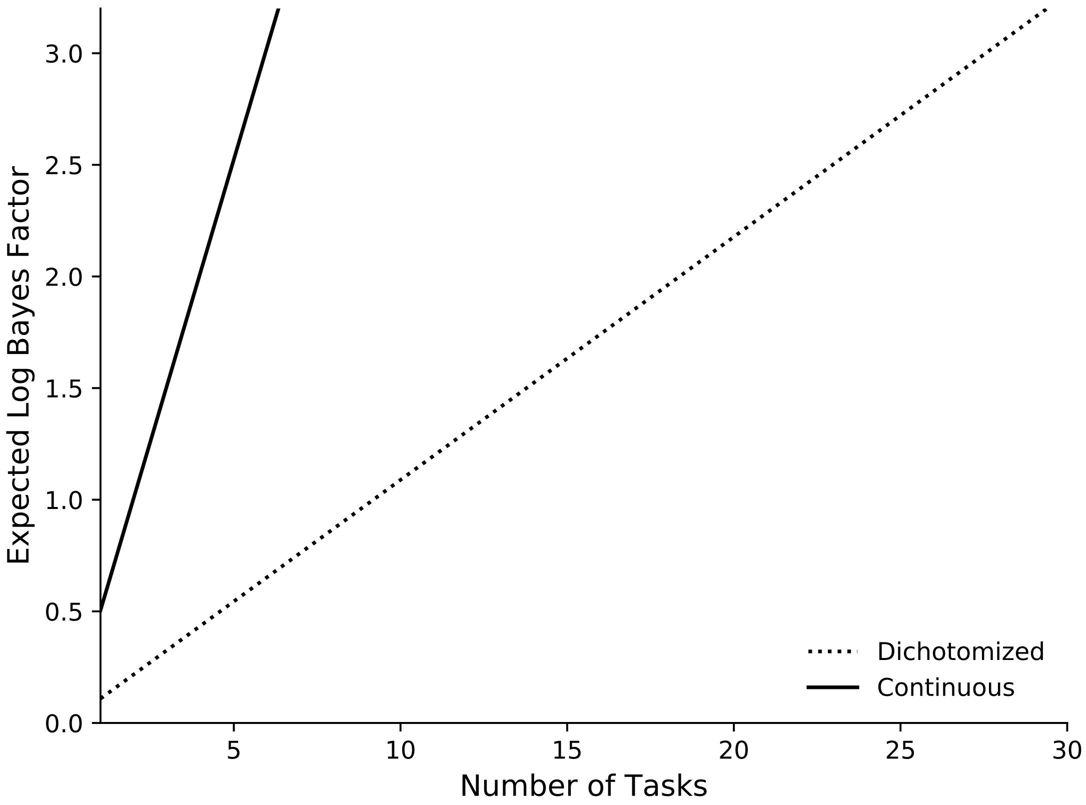


Figure A.1: Power analysis for environments with known feature ranking. The plot illustrates how many tasks are on average required to distinguish the ideal observer model from the single cue heuristic, assuming that decisions are made by the single cue heuristic. We show results for both dichotomized environments (dotted) and environments with continuous features (solid).

The expectation over tasks was approximated using 10^5 samples. Furthermore, we assumed that tasks are sampled independently from each other, meaning that we can multiply $\log B$ by the total number of encountered tasks K to get expected log-Bayes Factors for an experiment with K tasks. Figure A.1 shows this analysis for both continuous and dichotomized environments. We observed that it requires roughly four times more tasks to distinguish the single cue heuristic from an ideal observer model in environments with dichotomized features compared to one with continuous features.

B

Black-Box Variational Inference

Chapters 4 and 5 include models based on *black-box variational inference* (Ranganath et al., 2014). In both cases, initial priors are set to a standard normal distribution and variational posterior distributions are parametrized through a multivariate normal distribution, i.e. $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. The decision-making models of Chapter 4 adopt a mean field approximation, where posterior covariance matrices are restricted to be diagonal. To ensure positive semi-definite covariance matrices we parametrize them with logarithms of their standard deviations. The regression models of Chapter 5 parametrize the covariance matrix through a diagonal plus low rank factorization (Barber and Bishop, 1998) as described in Equation B.1. For all our model simulations we use a rank of $r = 1$.

$$\boldsymbol{\Sigma}_t = \operatorname{diag}\left(\exp(\boldsymbol{\upsilon}_t)\right)^2 + \mathbf{F}_t\mathbf{F}_t^T, \qquad \mathbf{F}_t \in \mathbb{R}^{d\times r} \tag{B.1}$$

Evidence lower bounds (ref. Equations 4.3 and 5.8) are optimized after each observation using the AmsGrad optimizer (Reddi et al., 2019). Training is stopped once the objective does not improve anymore over ten steps or after 1000 total gradient steps; the optimization procedure has typically converged at this point. The KL-divergence term is evaluated in closed-form, whereas we approximate the expected log-likelihood term through 100 samples and use the reparametrization trick (Kingma and Welling, 2013) to obtain gradients w.r.t. the variational parameters. The decision-making models use a learning rate of 0.1, whereas the regression models use a learning rate of 0.05.

C
Meta-Learning Details

BMI and RL^3 are obtained by minimizing Equation 4.16, Equation 5.9 and Equation 6.12 using the AmsGrad optimizer (Reddi et al., 2019). Learning rates are set to $3 \cdot 10^{-4}$ for the models from Chapter 4 and to 10^{-3} for the models from Chapters 5 & 6. Each model is initialized from a pretrained version without resource limitations and we increase β linearly to the desired value. We train for 10^6 iterations with a batch size of 32; at the end of meta-learning the loss function has converged.

Model architectures for all studies consists of a GRU with a hidden size of 128 units, which is followed by:

- a linear projection to a posterior distribution over probit regression weights (Chapter 4).
- a linear projection to the mean and log standard deviation of the predictive posterior distribution (Chapter 5).
- a linear projection to the mean of the optimal action-value function (Chapter 6). We construct a distribution over the optimal action-value function using this mean and a constant standard deviation of 10.

We employ the variational dropout prior in all models and parametrize the encoding distribution $q(\mathbf{\Theta}|\mathbf{\Lambda})$ through a fully factorized normal distribution (ref. Section 3.3.3). During meta-learning, the expectation of the log-likelihood term is approximated through one sample from the encoding distribution $q(\mathbf{\Theta}|\mathbf{\Lambda})$, and we obtain gradients with respect to $\boldsymbol{\Lambda}$ using the reparametrization trick (Kingma and Welling, 2013). During evaluation, the expectation of the log-likelihood term is approximated through $K = 100$ samples from the encoding distribution, and we perform no further updates of meta-parameters.

D
Bayesian Model Comparison

Bayesian model comparison (Bishop, 2006) provides us with a principled tool for comparing the evidence of different models. For the most part we perform separate comparisons for each participant and compute the probability that participant i used model m via Bayes' theorem:

$$p(m|\mathcal{D}_i) = \frac{p(\mathcal{D}_i|m)p(m)}{p(\mathcal{D}_i)} \tag{D.1}$$

The evidence for the decision-making models from Chapter 4 is given by:

$$\begin{aligned} p(\mathcal{D}_i|m) &= \prod_{k=1}^{K}\prod_{t=1}^{T} p(\mathcal{D}_{i,k,t}|m) && \text{(D.2)} \\ &= \prod_{k=1}^{K}\prod_{t=1}^{T} p(C_{i,k,t} = \hat{c}_{i,k,t}|\mathbf{x}_{i,k,t}, m) && \text{(D.3)} \end{aligned}$$

$\hat{c}_{i,k,t}$ denotes the decision made by participant i in task k and trial t and $\mathbf{x}_{i,k,t}$ denotes the corresponding input vector. K refers to the total number of tasks and T to the number of trials per task.

The evidence for the exploration models from Chapter 6 is given by:

$$p(\mathcal{D}_i|m) = \prod_{k=1}^{K}\prod_{t=1}^{T} p(\mathcal{D}_{i,k,t}|m) \tag{D.4}$$

$$= \prod_{k=1}^{K}\prod_{t=1}^{T} p(A_{i,k,t} = \hat{a}_{i,k,t}|m) \tag{D.5}$$

$\hat{a}_{i,k,t}$ denotes the action taken by participant i in task k and trial t.

In all our analyses, we assume a uniform prior over models. For some models, we additionally want to fit model parameters to empirical data. In these cases, we determine the parameter value that best describes each participant and approximate the model evidence using the *Bayesian information criterion* (BIC, Schwarz et al., 1978):

$$\log p(\mathcal{D}_i|m) \approx -\frac{1}{2}\log KT + \max_{\theta} \log p(\mathcal{D}_i|m, \theta) \tag{D.6}$$

Fitted parameters include the learning rate α of neural networks and the regularization factor β of BMI and RL^3. Finding the exact parameter value that maximize the model evidence for the given models is difficult, and thus we approximate it using a discrete set of candidate values. These are:

Chapter	**Model**	**Parameter**	**Values**
4	Neural network	α	$\{0, 2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}\}$
4	BMI	β	$\{0, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1\}$
6	RL^3	$\log_2 \beta$	$\{-13, -12, -11, -10, -9, -8, -7\}$

References

Joshua T Abbott and Thomas L Griffiths. Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

John R Anderson. The adaptive nature of human categorization. *Psychological review*, 98(3):409, 1991a.

John R. Anderson. Is human cognition adaptive? *Behavioral and Brain Sciences*, 14(3):471–485, 1991b. doi: 10.1017/S0140525X00070801.

F Gregory Ashby and W Todd Maddox. Human category learning. *Annu. Rev. Psychol.*, 56:149–178, 2005.

Anthony B Atkinson et al. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Shahar Ayal and GUY Hochman. Ignorance or integration: The cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making*, 22 (4):455–474, 2009.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.

David Barber and Christopher M Bishop. Ensemble learning for multi-layer networks. In *Advances in neural information processing systems*, pages 395–401, 1998.

Antoine Bechara, Antonio R Damasio, Hanna Damasio, Steven W Anderson, et al. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50:1–3, 1994.

Sara-Lee Been, Chris J Mitchell, Mark E Bouton, Russell Frohardt, et al. Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition*, 31(1):133–142, 2003.

Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *bioRxiv*, page 738385, 2020.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

Boris Belousov, Gerhard Neumann, Constantin A Rothkopf, and Jan R Peters. Catching heuristics are optimal control policies. In *Advances in neural information processing systems*, pages 1426–1434, 2016.

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule.* Citeseer, 1990.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

James O Berger. *Statistical decision theory and Bayesian analysis.* Springer Science & Business Media, 2013.

F Bryan Bergert and Robert M Nosofsky. A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1):107, 2007.

Ken Binmore. Rational decisions in large worlds. *Annales d'Economie et de Statistique*, pages 25–41, 2007.

Marcel Binz and Dominik Endres. Where do heuristics come from? In *CogSci*, pages 1402–1408, 2019.

Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.

Berndt Brehmer. Preliminaries to a psychology of inference. *Scandinavian Journal of Psychology*, 20(1):193–210, 1979.

Henry Brighton. Robust inference with simple cognitive models. In *AAAI spring symposium: Between a rock and a hard place: Cognitive science principles meet AI-hard problems*, pages 17–22, 2006.

Henry Brighton and Gerd Gigerenzer. Are rational actor models “rational” outside small worlds. *Evolution and Rationality: Decisions, Co-operation, and Strategic Behavior*, pages 84–109, 2012.

Arndt Bröder. Assessing the empirical validity of the” take-the-best” heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1332, 2000.

Arndt Bröder and Wolfgang Gaissmaier. Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review*, 14 (5):895–900, 2007.

Arndt Bröder and Stefanie Schiffer. Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General*, 132(2):277, 2003.

Arndt Bröder and Stefanie Schiffer. Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making*, 19 (4):361–380, 2006.

Egon Brunswik. *Perception and the representative design of psychological experiments.* Univ of California Press, 1956.

Eunhee Byun. *Interaction between prior knowledge and type of nonlinear relationship on function learning.* PhD thesis, ProQuest Information & Learning, 1996.

Colin Camerer, George Loewenstein, and Drazen Prelec. Neuroeconomics: How neuroscience can inform economics. *Journal of economic Literature*, 43(1):9–64, 2005.

Susan Carey and Elsa Bartlett. Acquiring a single new word. 1978.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems*, pages 5174–5185, 2019.

Gregory J Chaitin. On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM (JACM)*, 16(3):407–422, 1969.

Gregory J Chaitin. On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. *arXiv preprint math/0210035*, 2002.

Nick Chater and Paul Vitányi. Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1):19–22, 2003.

Nick Chater, Mike Oaksford, Ramin Nakisa, and Martin Redington. Fast, frugal, and rational: How rational norms explain behavior. *Organizational behavior and human decision processes*, 90(1):63–86, 2003.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Irene Cogliati Dezza, Axel Cleeremans, and William Alexander. Should we control? the interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, 148(6):977, 2019.

Aaron C Courville and Nathaniel D Daw. The rat as particle filter. In *Advances in neural information processing systems*, pages 369–376, 2008.

Aaron C Courville, Geoffrey J Gordon, David S Touretzky, and Nathaniel D Daw. Model uncertainty in classical conditioning. In *Advances in neural information processing systems*, pages 977–984, 2004.

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Similarity and discrimination in classical conditioning: A latent variable account. In *Advances in neural information processing systems*, pages 313–320, 2005.

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, 10(7): 294–300, 2006.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G Goldstein. How good are simple heuristics? In *Simple heuristics that make us smart*, pages 97–118. Oxford University Press, 1999.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Ishita Dasgupta, Eric Schulz, Joshua B Tenenbaum, and Samuel J Gershman. A theory of learning to infer. *Psychological Review*, 127(3):412, 2020.

Nathaniel D Daw, Aaron C Courville, and Peter Dayan. Semi-rational models of conditioning: The case of trial order. *The probabilistic mind*, pages 431–452, 2008.

Robyn M Dawes and Bernard Corrigan. Linear models in decision making. *Psychological bulletin*, 81(2):95, 1974.

Peter Dayan and Sham Kakade. Explaining away in weight space. In *Advances in neural information processing systems*, pages 451–457, 2001.

Peter Dayan and Theresa Long. Statistical models of conditioning. In *Advances in neural information processing systems*, pages 117–123, 1998.

Edward L DeLosh, Jerome R Busemeyer, and Mark A McDaniel. Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4):968, 1997.

Anja Dieckmann and Jörg Rieskamp. The influence of information redundancy on probabilistic inferences. *Memory & Cognition*, 35(7):1801–1813, 2007.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Michael O Duff. Optimal learning: Computational procedures for bayes-adaptive markov decision processes. 2003.

Michael O'Gordon Duff and Andrew Barto. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Hillel J Einhorn and Robin M Hogarth. Unit weighting schemes for decision making. *Organizational behavior and human performance*, 13(2):171–192, 1975.

Kevin Ellis, Armando Solar-Lezama, and Josh Tenenbaum. Sampling for bayesian program learning. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2016.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

Ido Erev and Greg Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912, 2005.

Jacob Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633, 2000.

Jacob Feldman. The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(5):330–340, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322, 2018.

Jerry A Fodor, Zenon W Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713, 2018.

Wilson S Geisler. Sequential ideal-observer analysis of visual discriminations. *Psychological review*, 96(2):267, 1989.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

Samuel J Gershman. A unifying probabilistic view of associative learning. *PLoS Comput Biol*, 11(11):e1004567, 2015.

Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

Samuel J Gershman. Uncertainty and exploration. *Decision*, 6(3):277, 2019.

Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 2020.

Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

Stefano Ghirlanda and Magnus Enquist. A century of generalization. *Animal Behaviour*, 66(1):15–36, 2003.

Soumya Ghosh and Finale Doshi-Velez. Model selection in bayesian neural networks via horseshoe priors. *arXiv preprint arXiv:1705.10388*, 2017.

Gerd Gigerenzer. On narrow norms and vague heuristics: A reply to kahneman and tversky. 1996.

Gerd Gigerenzer. The adaptive toolbox and lifespan development: Common questions? In *Understanding Human Development*, pages 423–435. Springer, 2003.

Gerd Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3 (1):20–29, 2008.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62:451–482, 2011.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Gerd Gigerenzer and Daniel G Goldstein. Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart*, pages 75–95. Oxford University Press, 1999.

Gerd Gigerenzer and Reinhard Selten. *Bounded rationality: The adaptive toolbox*. MIT press, 2002.

Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

Andreas Glöckner and Tilmann Betsch. Multiple-reason decision making based on automatic processing. *Journal of experimental psychology: Learning, memory, and cognition*, 34(5):1055, 2008.

Mark A Gluck and Gordon H Bower. From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117 (3):227, 1988.

Mark A Gluck, Daphna Shohamy, and Catherine Myers. How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory*, 9(6):408–418, 2002.

Daniel G Goldstein and Gerd Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive science*, 32 (1):108–154, 2008.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*, 2018.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016.

Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

Thomas L Griffiths, Edward Vul, and Adam N Sanborn. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268, 2012.

Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29: 24–30, 2019.

Peter D Grünwald and Abhijit Grunwald. *The minimum description length principle*. MIT press, 2007.

Kenneth R Hammond. Probabilistic functioning and the clinical method. *Psychological review*, 62(4):255, 1955.

Benjamin E Hilbig. Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, 17(6):923–930, 2010.

Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SklGryBtwr.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Mark K Ho, David Abel, Jonathan D Cohen, Michael L Littman, and Thomas L Griffiths. The efficiency of human cognition reflects planned information processing. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001a.

Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001b.

Robin M Hogarth and Natalia Karelaia. Ignoring information in binary choice with continuous variables: When is less “more”? *Journal of Mathematical Psychology*, 49(2):115–124, 2005.

Robin M Hogarth and Natalia Karelaia. Take-the-best and other simple strategies: Why and when they work well with binary cues. *Theory and Decision*, 61 (3):205–249, 2006.

Robin M Hogarth and Natalia Karelaia. Heuristic and linear models of judgment: Matching rules and environments. *Psychological review*, 114(3):733, 2007.

Antti Honkela and Harri Valpola. Variational learning and bits-back coding: an information-theoretic view to bayesian learning. *IEEE transactions on Neural Networks*, 15(4):800–810, 2004.

David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

John MC Hutchinson and Gerd Gigerenzer. Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural processes*, 69(2):97–124, 2005.

Quentin JM Huys, Tiago V Maia, and Michael J Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, 19(3):404, 2016.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

Jana B. Jarecki, Jolene H. Tan, and Mirjam A. Jenny. A framework for building cognitive process models. *Psychonomic Bulletin & Review*, Jul 2020. ISSN 1531-5320. doi: 10.3758/s13423-020-01747-2. URL https://doi.org/10.3758/s13423-020-01747-2.

Edwin T Jaynes. *Probability theory: The logic of science.* Cambridge university press, 2003.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Sharu Theresa Jose and Osvaldo Simeone. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization, 2020.

Peter Juslin, Sari Jones, Henrik Olsson, and Anders Winman. Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):924, 2003a.

Peter Juslin, Henrik Olsson, and Anna-Carin Olsson. Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, 132(1):133, 2003b.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600, 2012.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583, 2015.

Friso H Kingma, Pieter Abbeel, and Jonathan Ho. Bit-swap: Recursive bits-back coding for lossless compression with hierarchical latent variables. *arXiv preprint arXiv:1905.06845*, 2019.

David C Knill and Whitman Richards. *Perception as Bayesian inference.* Cambridge University Press, 1996.

Andrei N Kolmogorov. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):1–7, 1965.

Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

John K Kruschke. Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22, 1992.

John K Kruschke. Toward a unified model of attention in associative learning. *Journal of mathematical psychology*, 45(6):812–863, 2001.

John K Kruschke. Attention in learning. *Current Directions in Psychological Science*, 12(5):171–175, 2003.

John K Kruschke. Locally bayesian learning with applications to retrospective revaluation and highlighting. *Psychological review*, 113(4):677, 2006.

David A Lagnado, Ben R Newell, Steven Kahan, and David R Shanks. Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, 135(2):162, 2006.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR, 2018.

Brenden M Lake. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems*, pages 9791–9801, 2019.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350 (6266):1332–1338, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

Michael D Lee and Tarrant DR Cummins. Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic bulletin & review*, 11(2):343–352, 2004.

Timothy D Lee, Gabriele Wulf, and Richard A Schmidt. Contextual interference in motor learning: Dissociated effects due to the nature of task variations. *The Quarterly Journal of Experimental Psychology Section A*, 44(4):627–644, 1992.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

Stephan Lewandowsky and Simon Farrell. *Computational modeling in cognition: Principles and practice.* SAGE publications, 2010.

Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

Jan M Lichtenberg and Özgür Şimşek. Simple regression models. In *Imperfect Decision Makers: Admitting Real-World Rationality*, pages 13–25, 2017.

Falk Lieder and Thomas L Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762, 2017.

Falk Lieder and Thomas L Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 2020.

Falk Lieder, Paul M Krueger, and Tom Griffiths. An automatic method for discovering rational heuristics for risky choice. In *CogSci*, 2017.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*, 2017.

Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *Advances in neural information processing systems*, pages 3288–3298, 2017.

Shenghua Luan, Lael J Schooler, and Gerd Gigerenzer. From perception to preference and on to inference: An approach–avoidance analysis of thresholds. *Psychological Review*, 121(3):501, 2014.

Christopher G Lucas, Thomas L Griffiths, Joseph J Williams, and Michael L Kalish. A rational model of function learning. *Psychonomic bulletin & review*, 22(5):1193–1215, 2015.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Phil Maguire, Oisin Mulhall, Rebecca Maguire, and Jessica Taylor. Compressionism: a theory of mind based on data compression. In *CEUR Workshop Proceedings*, volume 1419, pages 294–299. CEUR, 2015.

Julian N Marewski and Lael J Schooler. Cognitive niches: an ecological model of strategy selection. *Psychological review*, 118(3):393, 2011.

Julian N Marewski, Wolfgang Gaissmaier, and Gerd Gigerenzer. Good judgments do not require complex cognition. *Cognitive processing*, 11(2):103–121, 2010.

Ellen M Markman. *Categorization and naming in children: Problems of induction.* Mit Press, 1989.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., USA, 1982. ISBN 0716715678.

Laura Martignon and Ulrich Hoffrage. Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52:29–71, 02 2002. doi: 10.1023/A:1015516217425.

David McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

James L. McClelland. The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38, 2009. doi: 10.1111/j.1756-8765.2008.01003.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2008.01003.x.

James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

James L McClelland, Matthew M Botvinick, David C Noelle, David C Plaut, Timothy T Rogers, Mark S Seidenberg, and Linda B Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

R Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L Griffiths, and Tal Linzen. Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*, 2020.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, art. arXiv:1802.03426, February 2018.

Björn Meder, Charley M Wu, Eric Schulz, and Azzurra Ruggeri. Development of directed and random exploration in children. 2020.

Douglas L Medin and Jeffrey G Bettger. Sensitivity to changes in base-rate information. *The American Journal of Psychology*, pages 311–332, 1991.

Vladimir Mikulik, Grégoire Delétang, Tom McGrath, Tim Genewein, Miljan Martic, Shane Legg, and Pedro A Ortega. Meta-trained agents implement bayes-optimal agents. *arXiv preprint arXiv:2010.11223*, 2020.

Ralph R Miller, Robert C Barnet, and Nicholas J Grahame. Assessment of the rescorla-wagner model. *Psychological bulletin*, 117(3):363, 1995.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2498–2507. JMLR. org, 2017.

Ben R Newell and Michael D Lee. The right tool for the job? comparing an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, 24(5):456–481, 2011.

Ben R Newell, Nicola J Weston, and David R Shanks. Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes*, 91(1):82–96, 2003.

Ben R Newell, David A Lagnado, and David R Shanks. Challenging the role of implicit processes in probabilistic category learning. *Psychonomic bulletin & review*, 14(3):505–511, 2007.

Ben R Newell, Nicola J Weston, Richard J Tunney, and David R Shanks. The effectiveness of feedback in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, 62(5):890–908, 2009.

Steven J Nowlan and Geoffrey E Hinton. Simplifying neural networks by soft weight-sharing. *Neural computation*, 4(4):473–493, 1992.

Mike Oaksford, Nick Chater, et al. *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford University Press, 2007.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4):481–487, 2004.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33, 2020.

Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.

Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Paula Parpart, Eric Schulz, Maarten Speekenbrink, and Bradley C. Love. Active learning reveals underlying decision strategies. *bioRxiv*, 2017. doi: 10.1101/239558. URL https://www.biorxiv.org/content/early/2017/12/25/239558.

Paula Parpart, Matt Jones, and Bradley C Love. Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, 102:127–144, 2018.

John W Payne, James R Bettman, and Eric J Johnson. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.

John W Payne, John William Payne, James R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge university press, 1993.

Steven T Piantadosi, Joshua B Tenenbaum, and Noah D Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123(4):392, 2016.

Neil C Rabinowitz. Meta-learners' learning dynamics are unlike learners'. *arXiv preprint arXiv:1905.01320*, 2019.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

Carl Edward Rasmussen and Zoubin Ghahramani. Occam's razor. In *Advances in neural information processing systems*, pages 294–300, 2001.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

RA Rescorla and Allan Wagner. *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, volume Vol. 2. 01 1972.

Jörg Rieskamp and Ulrich Hoffrage. When do people use simple heuristics, and how can we tell? 1999.

Jörg Rieskamp and Philipp E Otto. Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135(2):207, 2006.

Lionel Rigoux, Klaas Enno Stephan, Karl J Friston, and Jean Daunizeau. Bayesian model selection for group studies—revisited. *Neuroimage*, 84:971–985, 2014.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5): 465–471, 1978.

Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2940–2949. JMLR. org, 2017.

Jonas Rothfuss, Vincent Fortuin, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. *arXiv preprint arXiv:2002.05551*, 2020.

Joshua Rule, Eric Schulz, Steven T Piantadosi, and Joshua B Tenenbaum. Learning list concepts through program induction. *BioRxiv*, page 321505, 2018.

David E Rumelhart and James L McClelland. On learning the past tenses of english verbs. 1986.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.

Richard Samuels, Stephen Stich, and Michael Bishop. Ending the rationality wars. *Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010*, 2:191, 2012.

Adam N Sanborn and Ricardo Silva. Constraining bridges between levels of analysis: A computational justification for locally bayesian learning. *Journal of Mathematical Psychology*, 57(3-4):94–106, 2013.

Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144, 2010.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016.

Leonard J Savage. *The foundations of statistics.* Courier Corporation, 1972.

Benjamin Scheibehenne and Bettina Von Helversen. Useful heuristics. In *Making essential choices with scant information*, pages 195–212. Springer, 2009.

Benjamin Scheibehenne, Jörg Rieskamp, and Eric-Jan Wagenmakers. Testing adaptive toolbox models: A bayesian hierarchical approach. *Psychological review*, 120(1):39, 2013.

Juergen Schmidhuber, Jieyu Zhao, and MA Wiering. Simple principles of metalearning. *Technical report IDSIA*, 69:1–23, 1996.

E Schulz, M Speekenbrink, and B Meder. Simple trees in complex forests: Growing take the best by approximate bayesian computation. Cognitive Science Society, 2016a.

Eric Schulz and Samuel J Gershman. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55:7–14, 2019.

Eric Schulz, Josh Tenenbaum, David K Duvenaud, Maarten Speekenbrink, and Samuel J Gershman. Probing the compositionality of intuitive functions. In *Advances in neural information processing systems*, pages 3729–3737, 2016b.

Eric Schulz, Charley M Wu, Azzurra Ruggeri, and Björn Meder. Searching for rewards like a child means less generalization and more directed exploration. *Psychological science*, 30(11):1561–1572, 2019.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Dennis M Shaffer, Scott M Krauchunas, Marianna Eddy, and Michael K McBeath. How dogs navigate to catch frisbees. *Psychological Science*, 15(7): 437–441, 2004.

Anuj K Shah and Daniel M Oppenheimer. Heuristics made easy: An effort-reduction framework. *Psychological bulletin*, 134(2):207, 2008.

David R Shanks. Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, 37(1b): 1–21, 1985.

David R Shanks. *The psychology of associative learning*, volume 13. Cambridge University Press, 1995.

David R Shanks and Richard J Darby. Feature-and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 24(4):405, 1998.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Roger N Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4): 325–345, 1957.

Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

Hava T Siegelmann and Eduardo D Sontag. On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 440–449, 1992.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

Herbert A Simon. Bounded rationality. In *Utility and probability*, pages 15–18. Springer, 1990a.

Herbert A. Simon. Invariants of human behavior. *Annual Review of Psychology*, 41(1):1–20, 1990b. doi: 10.1146/annurev.ps.41.020190.000245. URL https://doi.org/10.1146/annurev.ps.41.020190.000245. PMID: 18331187.

Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.

Maarten Speekenbrink and Emmanouil Konstantinidis. Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, 7(2):351–367, 2015.

Kenneth O Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1): 24–35, 2019.

Mark Steyvers, Michael D Lee, and Eric-Jan Wagenmakers. A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, 2009.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Joshua B Tenenbaum and Thomas L Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, 24(4):629, 2001.

Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Peter M Todd and Anja Dieckmann. Heuristics for ordering cue search in decision making. In *Advances in neural information processing systems*, pages 1393–1400, 2005.

Peter M Todd and Gerd Gigerenzer. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(5):727–741, 2000.

Peter M Todd and Gerd Gigerenzer. Environments that make us smart: Ecological rationality. *Current directions in psychological science*, 16(3):167–171, 2007.

Peter M Todd and Gerd Ed Gigerenzer. *Ecological rationality: Intelligence in the world*. Oxford University Press, 2012.

James Townsend, Tom Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. *arXiv preprint arXiv:1901.04866*, 2019.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

Don Van Ravenzwaaij, Chris P Moore, Michael D Lee, and Ben R Newell. A hierarchical bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive Science*, 38(7):1384–1405, 2014.

Iris Van Rooij, Cory D Wright, and Todd Wareham. Intractability and the use of heuristics in psychological explanations. *Synthese*, 187(2):471–487, 2012.

Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive science*, 38 (4):599–637, 2014.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6): 860–868, 2018.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8 (3-4):279–292, 1992.

Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356, 1988.

Peter M Williams. Bayesian regularization and pruning using a laplace prior. *Neural computation*, 7(1):117–143, 1995.

Ian H Witten, Radford M Neal, and John G Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Charley M Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, 2(12):915–924, 2018.

Timo C Wunderlich and Christian Pehle. Eventprop: Backpropagation for exact gradients in spiking neural networks. *arXiv preprint arXiv:2009.08378*, 2020.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*, 2019.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

Carlos Zednik and Frank Jäkel. Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, 193(12):3951–3985, 2016.

Shunan Zhang and J Yu Angela. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems*, pages 2607–2615, 2013.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019a.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019b.