Title

Where Do Heuristics Come From?

Permalink

https://escholarship.org/uc/item/1gd685wb

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 41(0)

Authors

Binz, Marcel
Endres, Dominik

Publication Date

2019

Peer reviewed

Where Do Heuristics Come From?

Marcel Binz (binz@staff.uni-marburg.de)
Department of Psychology, Theoretical Neuroscience Group
Philipps-Universität Marburg

Dominik Endres (dominik.endres@uni-marburg.de)
Department of Psychology, Theoretical Neuroscience Group
Philipps-Universität Marburg

Abstract

Human decision-making deviates from the optimal solution, i.e. the one maximizing cumulative rewards, in many situations. Here we approach this discrepancy from the perspective of computational rationality and our goal is to provide justification for such seemingly sub-optimal strategies. More specifically we investigate the hypothesis, that humans do not know optimal decision-making algorithms in advance, but instead employ a learned, resource-constrained approximation. The idea is formalized through combining a recently proposed meta-learning model based on Recurrent Neural Networks with a resource-rational objective. The resulting approach is closely connected to variational inference and the Minimum Description Length principle. Empirical evidence is obtained from a two-armed bandit task. Here we observe patterns in our family of models that resemble differences between individual human participants.

Keywords: Bounded rationality; computational rationality; variational inference; reinforcement learning; meta-learning; individual differences; multi-armed bandit

Introduction

In this work we study human decision-making strategies on a stationary multi-armed bandit task. These are among the simplest sequential decision-making problems, that require reasoning about trade-offs between exploration and exploitation. In the special case of an infinite horizon and geometric discounting their Bayes-optimal solution is the Gittins index strategy (Gittins, 1979), while in general it is defined as the result of a planning process in an augmented Markov Decision Process (Duff & Barto, 2002). Prior work however suggests, that several heuristics appear to be favourable as a model of human decision-making, when compared to the Bayes-optimal solution (Steyvers, Lee, & Wagenmakers, 2009; Zhang & Angela, 2013).

Understanding human cognition in terms of heuristics has been a major theme in cognitive science over the past decades (Tversky & Kahneman, 1974; Simon, 1990; Gigerenzer & Todd, 1999). They can be viewed as crude, but realizable, approximations of optimal behavior. Heuristics are thus connected to the idea of rationality under resource constraints, which is commonly referred to as bounded rationality (Simon, 1972), computational rationality (Gershman, Horvitz, & Tenenbaum, 2015), or resource-rationality (Griffiths, Lieder, & Goodman, 2015). Examples for resource constraints include related prior experience on a given task, limited capacity of our brain or restricted deliberation times. For a more general overview of computational rationality we refer the reader to Gershman et al. (2015). Here we are interested in the hypothesis, that humans employ a learned, resourced-constrained approximation of an optimal decision-making strategy. More specifically we show, that different, potentially sub-optimal, human strategies emerge naturally in artificial learning systems when varying the strength of the constraints placed upon them. For a realization of this principle, we rely on information-theoretic concepts, similar to the approach of Ortega and Braun (2013).

We instantiate a particular kind of such resource-rational agents using recent advances from the meta-learning literature (Wang et al., 2016; Duan et al., 2016). In this framework the algorithm to be learned is parametrized by a Recurrent Neural Network (RNN). RNNs are known to be Turing-complete and hence are in theory able to realize any algorithm (Siegelmann & Sontag, 1991). The RNN is trained on a set of related tasks to act as an independent Reinforcement Learning algorithm for solving the original problem. We treat all parameters of the RNN as random variables and infer approximate posterior distributions by solving a regularized optimization problem. Varying the regularization factor leads to a spectrum of resource-rational algorithms, each possessing different properties. Models with large constraints need to rely more on prior assumptions and thus prefer simple strategies, while models with weaker constraints will approach the optimal solution (up to the representational capabilities of the RNN and the limitations of the meta-learning procedure).

The resulting approach is closely related to the Minimum Description Length (MDL) principle (Hinton & Van Camp, 1993; Grunwald, 2004), which asserts that the best model is the one, that leads to the best compression of the data, including a cost for describing the model. The bits-back argument establishes a link between the MDL principle and Bayesian learning (Honkela & Valpola, 2004), opening up connections to Bayesian theories of cognition (Griffiths, Kemp, & Tenenbaum, 2008). Indeed several heuristics have been recently interpreted as Bayesian models under strong priors (Parpart, Jones, & Love, 2018).

Our hypothesis is validated on a classical two-armed bandit task. However we view multi-armed bandits merely as

the first step towards investigating more complex tasks and the proposed algorithm is not limited to any specific problem class. The following section first introduces the framework in more general terms, before considering multi-armed bandits as a special case. We then identify different strategies of human participants and subsequently show how the proposed class of models captures important characteristics of human behavior on both a qualitative and quantitative level. Our results indicate, that the seemingly sub-optimal decision strategies used by humans might be a consequence of the constraints under which these very strategies are learned.

Methods

Reinforcement Learning

Let $M = (\mathcal{S}, \mathcal{A}, p, \gamma)$ be a Markov Decision Process (MDP), with a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a joint distribution over the next state and a scalar reward signal, describing the dynamics of the environment, $p(s_{t+1}, r_t | s_t, a_t)$ and a discount factor $\gamma \in [0, 1]$. The objective of a Reinforcement Learning (RL) agent is to find a policy $\pi(a_t | \cdot)$, that maximizes the discounted, expected return $\mathbb{E}_{p,\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$ without having direct access to the true underlying dynamics p.

Learning Reinforcement Learning Algorithms

Following the approach of Wang et al. (2016); Duan et al. (2016) we want to *learn* a RL algorithm for solving a MDP sampled from a distribution over MDPs. We parametrize the algorithm to be learned with a Recurrent Neural Network (RNN), in form of a Gated Recurrent Unit (Cho et al., 2014), followed by a linear layer. The set of all model parameters is denoted with θ in the following. The RNN takes previous actions and rewards as inputs in addition to the current state, making the output a function of the entire history $X_t = (s_0, a_0, r_0, s_1 \ldots, a_{t-1}, r_{t-1}, s_t)$. A good algorithm has to integrate information from the history in order to identify the currently active MDP, based on which it subsequently has to select the appropriate strategy. The RNN is trained to accomplish this using standard model-free RL techniques. In this work we utilize *n*-step Q-Learning (Mnih et al., 2016), although in theory any other algorithm could be applied as well. The RNN implements a freestanding RL algorithm through its recurrent activations after training is completed (the parameters of the RNN are held constant during evaluation). Throughout this work we use the abbreviation LRLA – for learned Reinforcement Learning algorithm – to refer to this kind of model. Alternatively we can view this procedure as a model-free algorithm for partially observable MDPs, where the hidden information consists of the currently active task.

Resource-Rational Decision-Making

We consider maximizing the following regularized objective for inferring a distribution q_ϕ over parameters θ of LRLAs:

$$\mathcal{L}(\phi, \mathbf{X}, \mathbf{y}) = \mathbb{E}_{q_\phi(\theta)} [\log p(\mathbf{y} | \mathbf{X}, \theta)] - \beta \mathrm{KL}(q_\phi(\theta) || p(\theta)) \quad (1)$$

where the hyperparameter β controls how much the posterior is allowed to deviate from the prior in terms of the

Kullback-Leibler (KL) divergence. We assume a likelihood $p(\mathbf{y} | \mathbf{X}, \theta)$, that factorizes over data points $\prod_{i=1}^{N} p(y_i | X_i, \theta)$ and we approximate each factor with a normal distribution of fixed scale σ_y: $\mathcal{N}(y_t; Q_\theta(X_t, a), \sigma_y)$. In our setting $Q_\theta(X_t, a)$ corresponds to the RNN output after seeing history X_t and y_t corresponds to the n-step return $\sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n \max_a Q_\theta(X_{t+n}, a)$. The corresponding policy is derived as follows:

$$\pi(a_t | X_t) = \begin{cases} 1 & \text{if } a_t = \arg\max_{a \in \mathcal{A}} Q_\theta(X_t, a) \\ 0 & \text{else} \end{cases} \quad (2)$$

Setting β to a specific value can be interpreted as implicitly defining a constraint on $\mathrm{KL}(q_\phi(\theta) || p(\theta))$. Importantly the KL term determines how much the model parameters can be compressed in theory (Hinton & Van Camp, 1993). Hence our models are resource-constrained with regard to a hypothetical lower bound on their storage capacity. Intuitively, if the regularization factor β is large, parameters are forced to match the prior closely. In this work we employ priors favoring simple functions, hence models are only allowed to realize more complex functions as $\beta \to 0$.

Bayesian Interpretation

If we set $\beta = 1$, we recover the evidence lower bound (ELBO) as an objective for performing variational inference. In the setting of large data-sets subsampling techniques are often employed to approximate Equation 1 using mini-batches $\mathcal{B}$ of size M with an appropriately scaled log-likelihood term:

$$\log p(\mathbf{y} | \mathbf{X}, \theta) \approx \frac{N}{M} \sum_{i \in \mathcal{B}} \log p(y_i | X_i, \theta) \quad (3)$$

If data arrives in sequential fashion, as it does in the RL setting, the data-set size N is not known in advance and has to be treated as an additional hyperparameter. This leads to a Bayesian interpretation of Equation 1 even for $\beta \neq 1$. For any values of β and N maximizing Equation 1 is equivalent to performing stochastic variational inference with an assumed data-set size of $\hat{N} = \frac{N}{\beta}$. In practice we optimize a by N^{-1} scaled version of Equation 1, which leads to $\hat{N}^{-1}$ as a factor for the KL term.

In the following section we investigate whether we can understand individual differences in human decision-making in terms of optimal solutions to Equation 1 for varying values of β. It is worth clarifying, that we are only interested in the computational aspects of this hypothesis, i.e. we want to test, whether human decision-making can be characterized through resource-rational strategies. We do not attempt to answer how this objective is realized on an algorithmic or implementational level.

Technical Details

We maximize Equation 1 using standard gradient-based optimization techniques. For this we simulate k environments in
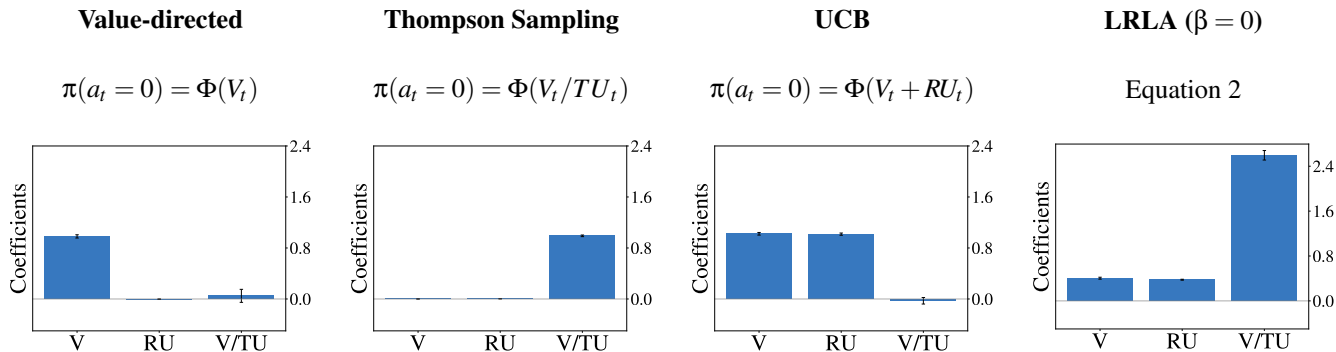
Value-directed	Thompson Sampling	UCB	LRLA ($\beta = 0$)
$\pi(a_t = 0) = \Phi(V_t)$	$\pi(a_t = 0) = \Phi(V_t/TU_t)$	$\pi(a_t = 0) = \Phi(V_t + RU_t)$	Equation 2

Figure 1: Illustration of different algorithms for two-armed bandits. **Middle**: Definitions of the respective policy. **Bottom**: Coefficients obtained from fitting the probit regression (Equation 5) to corresponding trajectories. Error bars indicate the uncertainty (one standard deviation) in the coefficients estimated through a Laplace approximation. Note, that for LRLAs the coefficients are task-dependent. For this plot we use the set of two-armed bandits described in the later sections to compute the coefficients. Φ denotes the cumulative distribution function of a standard normal distribution.

parallel and update the model at the end of each episode. All models in this work employ a group horseshoe prior, which can be viewed as a continuous relaxation of a spike-and-slab prior (Mitchell & Beauchamp, 1988), over their weights:

$$s \sim C^+(0, \tau_0); \quad \tilde{z}_i \sim C^+(0, 1);$$
$$\tilde{\theta}_{ij} \sim \mathcal{N}(0, 1); \quad \theta_{ij} = \tilde{\theta}_{ij} \tilde{z}_i s$$

and we represent the approximate posterior $q_\phi(\theta)$ through a fully factorized distribution as proposed in (Louizos, Ullrich, & Welling, 2017). The hyperparameter of the horseshoe prior is fixed to $\tau_0 = 10^{-5}$. The horseshoe prior is a sparsity-inducing prior, which causes our models to implement simple functions in absence of any experience. During training we approximate the expectation of the log-likelihood term with a single sample from $q_\phi(\theta)$ and make use of the reparametrization trick (Kingma & Welling, 2013). Resampling of weight matrices is done only at the beginning of an episode as proposed in Gal and Ghahramani (2016); Fortunato, Blundell, and Vinyals (2017). Target values y_t are computed using the maximum a posteriori estimate of a separate target network (Mnih et al., 2013; Lipton et al., 2017). For additional details we refer the reader to the publicly available implementation[1].

Multi-Armed Bandits

Experiments in the following section involve a multi-armed bandit task. These are MDPs consisting of a single state. At each step t an agent selects one out of multiple actions and is rewarded according to an unknown, stationary distribution based on its choice. This interaction is repeated T times.

The trade-off between exploiting good options and exploring yet unknown ones is the central theme in multi-armed bandits (and in RL in general). Methods for resolving this exploration-exploitation dilemma can be categorized in 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. For this we consider a two-armed bandit task with normal distributions over both the mean of rewards for each arm and their reward noise at each time-step. Let $\mathcal{N}(r_a; \mu_{0,a}, \sigma_{0,a})$ be an independent normal prior over expected rewards for each action a and $\mathcal{N}(r_a; \mu_{t,a}, \sigma_{t,a})$ be the posterior after t interactions. Many popular strategies can be formulated using the parameters of these distributions. Define:

$$V_t = \mu_{t,0} - \mu_{t,1}$$
$$RU_t = \sigma_{t,0} - \sigma_{t,1} \qquad (4)$$
$$TU_t = \sqrt{\sigma_{t,0}^2 + \sigma_{t,1}^2}$$

V_t constitutes the estimated difference in value, while RU_t and TU_t describe relative and total uncertainty respectively. Choice probability in Thompson sampling (an example for random exploration) is only a function of V_t and TU_t, while it is a function of V_t and RU_t for the UCB algorithm (an example of directed exploration). Figure 1 (middle row) shows definitions of all strategies under consideration. For a given set of observed trajectories $\mathcal{D}$ one can fit a probit regression model to infer the importance of factors from Equation 4:

$$p(a_t = 0 | \mathcal{D}, \mathbf{w}) = \Phi(w_1 V_t + w_2 RU_t + w_3 V_t / TU_t) \quad (5)$$

Analyzing the resulting coefficients $\mathbf{w}$ can reveal, which exploration strategy generated the observations, as shown in Figure 1 (bottom row). We utilize this form of analysis throughout the following sections.

Empirical Analysis

Human Participants

We initially inspect human exploration strategies on a two-armed bandit task with episode length $T = 10$. The mean

---

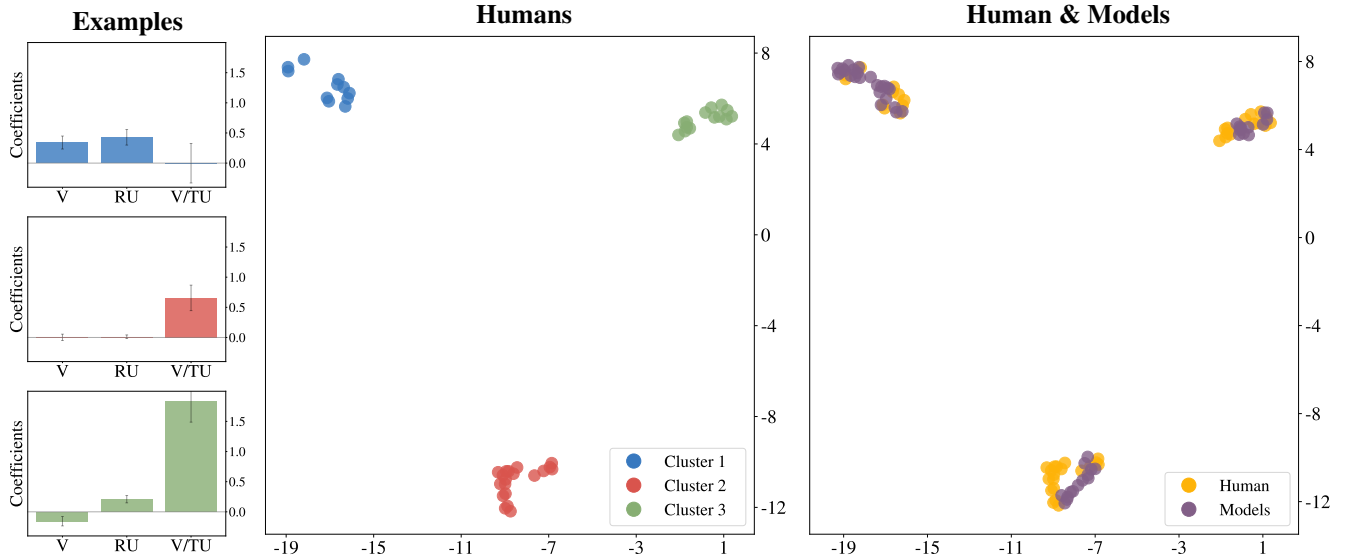[1] https://github.com/marcelbinz/MDLDQN

Figure 2: Visualization of human policies alongside resource-constrained LRLAs. **Left**: Probit regression coefficients of prototype participants. Prototypes were obtained from a mean-shift clustering, shown in the middle plot. Colors correspond to clusters. Error bars indicate the uncertainty (one standard deviation) in the coefficients estimated through a Laplace approximation. **Middle**: UMAP (McInnes & Healy, 2018) embedding of coefficients for all participants. **Right**: Joint UMAP embedding of coefficients for human participants and LRLAs $\in \mathcal{H}_{\mathrm{LRLA}}$.

reward for each action is drawn from $\mathcal{N}(\mu_a; 0, \sqrt{100})$ at the beginning of an episode and the reward in each step from $\mathcal{N}(r_t; \mu_{a_t}, \sqrt{10})$. Intuitively we expect some participants to be more proficient at the task, for example because they have more experience at related problems (higher $\hat{N}$), while the opposite is true for others. We rely on data gathered by Gershman (2018), which contains records of 44 participants, each playing 20 of the aforementioned two-armed bandit problems. Figure 2 (middle) shows the result of fitted probit regression coefficients for individual participants. This analysis reveals three major subgroups within the population, each using a different set of strategies. We visualize coefficients of three example participants (Figure 2, left) and observe, that a large fraction is well-described through Thompson sampling (clusters 2 and 3), while other participants have tendencies towards a mixture of strategies (cluster 1).

Learned Reinforcement Learning Algorithms

Next we show, that optimizing LRLAs with different regularization factors leads to the emergence of diverse exploration pattern. We train otherwise identical models for $\hat{N} \in \mathcal{H}_{\mathrm{LRLA}} = \{256, 512, 1024, 2048, 4096, 8192\}$ on the same two-armed bandit task until convergence and report average results over 10 random seeds unless otherwise noted. Equation 1 is approximated with a batch of samples from complete episodes of 16 parallel simulations and gradient-based optimization is performed using Adam (Kingma & Ba, 2014). Figure 3 (left) shows, that performances continuously improves as $\hat{N}$ increases, confirming our expectation that models become more sophisticated for large $\hat{N}$. Fitting the aforementioned probit regression model to the resulting policies (Fig-

ure 3, right) reveals value-based characteristics at one end of the spectrum. Towards the other end we observe coefficients, that slowly transition to those of the unconstrained ($\beta = 0$) model.

Modelling Human Behavior

We are mainly interested in whether the set of resource-constrained LRLAs can help us to understand human behavior on an individual level. To answer this question, we compare the optimized models to human decision-making strategies in terms of the probit regression analysis. We visualize the regression coefficients for 50 models (10 for each value of $\hat{N} \in \mathcal{H}_{\mathrm{LRLA}}$, excluding $\hat{N} = 256$) alongside those of the human participants in Figure 2 (right). Although some parts of the low-dimensional embedding are over- and underrepresented, the overall variation of human exploration strategies is captured by the resource-constrained LRLAs.

Model Comparison

The regression analysis performed so far provides only qualitative indicators for our hypothesis. In order to obtain a quantitative measure for the explanatory power of the proposed hypothesis, we performed a Bayesian model comparison. Figure 4 (left) shows log-likelihoods for each participant and model. We observe, that different participants are modelled best with different values of $\hat{N}$.

To verify that the class of resource-constrained LRLAs $\mathcal{H}_{\mathrm{LRLA}}$ contains a good model, we compute Bayes factors (BF) between the marginal probability of the resource-
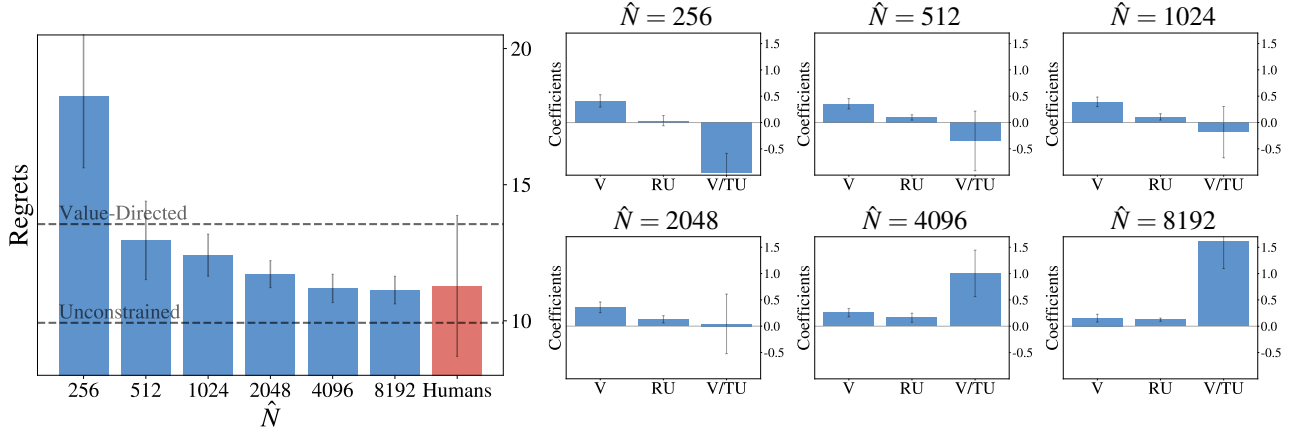
Figure 3: Results for optimized LRLAs with different $\hat{N}$. **Left**: Visualization of per episode regret averaged over 10 models and 1000 episodes. Horizontal lines correspond to the performance of a value-directed policy and an unconstrained LRLA. **Right**: Coefficients of the probit regression from Equation 5. Error bars indicate standard deviations across the 10 models.

constrained LRLAs and a value-directed policy:

$$\log BF_i = \log p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}}) - \log p(\mathcal{D}_i|H_{\text{value-directed}})$$

$$p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}}) = \frac{1}{|\mathcal{H}_{\text{LRLA}}|} \sum_{H \in \mathcal{H}_{\text{LRLA}}} p(\mathcal{D}_i|H) \tag{6}$$

where $\mathcal{D}_i$ refers to all actions taken by a specific participant and $\frac{1}{|\mathcal{H}_{\text{LRLA}}|}$ is a prior that corrects for multiple comparisons across different values of $\hat{N}$. The resulting $\log BF$s (see Figure 4, right) reveal strong evidence for 42 of the 44 participants in favor of the class of resource-constrained LRLAs, when compared with the baseline. This indicates, that one of the models in $\mathcal{H}_{\text{LRLA}}$ explains the participant's behavior much better than the value-directed policy. There are nine participants best described by letting $\hat{N} = 512$, nine by $\hat{N} = 1024$, 20 by $\hat{N} = 4096$ and six by $\hat{N} = 8192$. This heterogeneity highlights, that the model class is able to accomodate individual differences between human participants.

Finally we want to show, that the proposed class of models captures exploration strategies across all participants better than any standard exploration strategy alone. To verify this, we computed Bayes factors between $\prod_i p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}})$ and two baseline exploration strategies: $\prod_i p(\mathcal{D}_i|H_{\text{Thompson}})$ and $\prod_i p(\mathcal{D}_i|H_{\text{UCB}})$. We find $2\log BF = 72.8$ against Thompson sampling and 5391.4 against UCB, indicating that our class of models is overall better at representing exploration strategies for all participants in comparison to any single, fixed strategy.

Discussion

In this work we proposed a justification for seemingly sub-optimal human strategies in sequential decision-making problems based on the idea of computational rationality. We view human decision-making as an instance of a learned, resource-constrained RL algorithm. This is formalized through learning distributions over parameters of a meta-learning model with a regularized, resource-rational objective. The emerging spectrum of strategies resembles characteristics of human decision-making without being explicitly trained to do so. Additional model comparison suggests, that the resulting resource-constrained LRLAs describe human policies well on a quantitative level. However, the correspondence between human behavior and the LRLA model class is not perfect. Looking at Figure 2 (right) we observe, that some clusters are not represented exactly. Furthermore it remains open, why none of the participants is best described through the model with $\hat{N} = 2048$. Accounting for these observations remains a question for future work.

The analysis on the two-armed bandit task presented in this work can be extended in several ways. Relating deliberation times to regularization factors could, for example, provide additional evidence for our hypothesis. It also remains to be seen whether our conclusions transfer to other sequential decision-making problems beyond the bandit setting. In this context we are especially interested in tasks, where descriptive models of individual human behavior consist of a set of different heuristics. We are also interested in methods, that allow us to disentangle resource-rational behavior from the Bayesian interpretation.

Recent work on model-free meta-learning methods, similar to the one employed in this work, indicates an emergence of model-based behavior (Wang et al., 2016) and causal reasoning (Dasgupta et al., 2019) as well as the ability for few-shot learning (Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016), properties supposedly absent in artificial systems. Having systems capable of such feats, opens the possibility for interesting studies on human cognition.
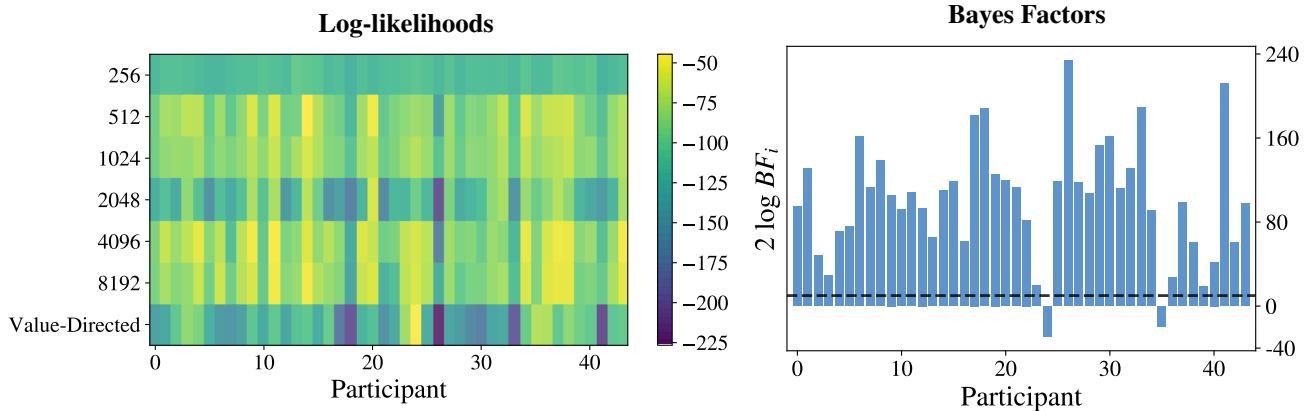
Figure 4: Model comparison of the set of resource-constrained LRLAs with a value-directed baseline. **Left**: Log-likelihoods for each participant and model. Higher values indicate a better fit. **Right**: Bayes factors (see Equation 6) for each participant i. The dotted horizontal line (equal to 10) corresponds to the threshold for very strong evidence (Kass & Raftery, 1995) in favour of $\mathcal{H}_{\text{LRLA}}$.

Acknowledgments

References

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., ... Kurth-Nelson, Z. (2019). *Causal reasoning from meta-reinforcement learning*. Retrieved from https://openreview.net/forum?id=H1ltQ3R9KQ

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Duff, M. O., & Barto, A. (2002). *Optimal learning: Computational procedures for bayes-adaptive markov decision processes*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Fortunato, M., Blundell, C., & Vinyals, O. (2017). Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems* (pp. 1019–1027).

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

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

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Evolution and Cognition (Paper.

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

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

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

Grunwald, P. (2004). A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*.

Hinton, G. E., & Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on computational learning theory* (pp. 5–13).

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lipton, Z., Li, X., Gao, J., Li, L., Ahmed, F., & Deng, L. (2017). Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*.

Louizos, C., Ullrich, K., & Welling, M. (2017). Bayesian compression for deep learning. In *Advances in neural information processing systems* (pp. 3288–3298).

McInnes, L., & Healy, J. (2018, February). UMAP: Uniform Manifold Approximation and Projection for Dimen-

sion Reduction. *ArXiv e-prints*.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A*, *469*(2153), 20120683.

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, *102*, 127–144.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, *4*(6), 77–80.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).