01. Marcel Binz, Ishita Dasgupta, Akshay Jagadish, Matthew Botvinick, Jane Wang, and Eric Schulz

02. Abstract word count: **60**; Main text word count: **990**; References word count: **438**; Entire text (total + addresses etc.) word count: **1577**

03. Combining meta-learned models with process models of cognition

04. Adam N. Sanborn; Haijiang Yan; Christian Tsvetkov

05. University of Warwick; University of Warwick; University of Warwick

06. University of Warwick, CV4 7AL, UK

07. Sanborn: +44 24 761 51354

08. a.n.sanborn@warwick.ac.uk; haijiang.yan@warwick.ac.uk; chris.tsvetkov@warwick.ac.uk

09. https://go.warwick.ac.uk/adamsanborn; na; na

10. Meta-learned models of cognition make optimal predictions for the actual stimuli presented to participants, but investigating judgment biases by constraining neural networks will be unwieldy. We suggest combining them with cognitive process models, which are more intuitive and explain biases. Rational process models, those that can sequentially sample from the posterior distributions produced by meta-learned models, seem a natural fit.

11. Meta-learned models of cognition offer an exciting opportunity to address a central weakness of current cognitive models, whether Bayesian or not: cognitive models generally do not “see” the experimental stimuli shown to participants. Experimenters instead feed models low-dimensional descriptions of the stimuli, which are often in terms of the psychological features imagined by the experimenter, or sometimes are the psychological descriptions that best fit participants’ judgments (e.g., stimulus similarity judgments; Nosofsky, et al., 2018).

For example, in studies of probability judgment, participants have been asked to judge the probability that “Bill plays jazz for a hobby” after having been given the description, “Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities” (Tversky & Kahneman, 1983). Current probability judgment models reduce

these descriptions down to a single unknown number, and attempt to find the latent probability that best fits the data (e.g., Zhu, et al., 2020).

Models trained on the underlying statistics of the environment, as meta-learned models are, can bypass this need to infer a latent variable, instead making predictions from the actual descriptions used. Indeed, even relatively simple models of semantics that locate phrases in a vector space produce judgments that correlate with the probabilities experimental participants give (Bhatia, 2017). Meta-learned models could thus explain a great deal of the variability in human behavior, and allow experimenters to generalize beyond the stimuli shown to participants.

However, used as descriptive models, normative meta-learned models of cognition inherit a fundamental problem from the Bayesian approach: people's reliable deviations from normative behavior. One compelling line of research shows that probability judgments are incoherent in a way that Bayesian models are not. Using the above example of Bill, Tversky and Kahneman (1983) found participants ranked the probability of "Bill is an accountant who plays jazz for a hobby" as higher than that of "Bill plays jazz for a hobby". This violates the extension rule of probability because the set of all accountants who play jazz for a hobby is a subset of all people who play jazz for a hobby, no matter how Bill is described.

The target article discusses constraining meta-learned models to better describe behavior, such as reducing the number of hidden units or restricting the representational fidelity of units. These manipulations have produced a surprising and interesting range of biases, including stochastic and incoherent probability judgments (Dasgupta, et al., 2020). However, this is just the start to explaining human biases. Even a single bias such as the conjunction fallacy has intricacies, such as the higher rate of conjunction fallacies when choosing versus estimating (Wedell & Moro, 2008), and greater variability in judgments of conjunctions than those of simple events (Costello & Watts, 2017).

Cognitive process models aim to explain these biases in detail. For conjunction fallacies, a variety of well-supported models exist, based on ideas such as participants sampling events with noise in the retrieval process (Costello & Watts, 2014), or by sacrificing probabilistic coherence to improve judgment accuracy based on samples (Zhu, et al., 2020), or by representing conjunctions as a weighted average of simple events (Juslin, et al., 2009), or by using quantum probability (Busemeyer, et al., 2011). These kinds of models capture many details of the empirical effects, through simple and intuitive mechanisms like adjusting the amount of noise or number of samples, which helps identify experiments to distinguish between them.

Mechanistically modifying meta-learned models to explain cognitive biases to the level cognitive process models do appears difficult. While changes to network structure are powerful ways to induce different biases that could identify implementation-level constraints in the brain, the effects of these kinds of changes are generally hard to intuit, while training constrained meta-learning models to test different manipulations will be slow and computationally expensive. Thus, it will be challenging to reproduce existing biases in detail or to design effective experiments for testing these constraints.

Combining meta-learned models with cognitive process models is more promising. One possibility is to have meta-learned models act as a “front end” that takes stimuli and converts them to a feature-based representation, which is then operated on by a cognitive process model. The parameters of the cognitive process model could be fit to human data, or potentially the cognitive process model could be encoded into the network (e.g., Peterson et al., 2021), and meta-learning could be done on the front end and the cognitive process parameters end-to-end.

However, as meta-learned models of cognition produce posterior predictive distributions, rational process models offer a straightforward connection that does not require retraining meta-learned models. Rational process models do not directly use a posterior predictive distribution, but instead assume that the posterior predictive distribution is approximated (i.e., using the posterior mean, posterior median, or other summary statistic depending on task), most often using a statistical sampling algorithm (Griffiths, et al., 2012). Such a model can explain details of the conjunction fallacy, and also a wide range of other biases, such stochastic choice, anchoring and repulsion effects in estimates, long-range autocorrelations in judgment, and the flaws in random sequence generation (Castillo, et al., 2024; Spicer, et al., 2022; Vul et al., 2014; Zhu et al., 2022; 2023). What these models have lacked, however, is a principled way in which to construct the posterior predictive distribution from environmental statistics, and here meta-learned models offer that exciting possibility.

While rational process models offer what we think is a natural choice for integration, any sort of combination with existing cognitive models offers benefits. Being able to explain both the details of biases as cognitive process models do, as well as showing sensitivity to actual stimuli is a powerful combination that moves toward the long-standing goal of a general model of cognition. Overall we see meta-learned models of cognition as not supplanting existing cognitive models, but as a way to make them much more powerful and relevant to understanding and predicting behavior.

12. Acknowledgments: none.

13. Competing interests: none

14. ANS and CT were supported by a European Research Council consolidator grant (817492- SAMPLING). HY was supported by a Chancellor's International Scholarship from the University of Warwick.

15. Alphabetical reference list

Bhatia, S. (2017) Associative judgment and vector space semantics. *Psychological Review, 124(1),* 1-20. http://dx.doi.org/10.1037/rev0000047

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review, 118*(2), 193–218. https://doi.org/10.1037/a0022542

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. (2024). Explaining the flaws in human random generation as local sampling with momentum. *PLOS Computational Biology*, *20*(1), e1011739. https://doi.org/10.1371/journal.pcbi.1011739

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review, 121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304-321. https://dx.doi.org/10.1002/bdm.1936

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review, 127*(3), 412–441. https://doi.org/10.1037/rev0000178

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263-268. https://doi.org/10.1177/0963721412447619

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*(4), 856–874. https://doi.org/10.1037/a0016979

Nosofsky, R.M., Sanders, C.A., Meagher, B.J. &, Douglas, B.J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods, 50,* 530–556. https://doi.org/10.3758/s13428-017-0884-8

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209-1214. https://doi.org/10.1126/science.abe2629

Spicer, J., Zhu, J. Q., Chater, N., & Sanborn, A. N. (2022). Perceptual and cognitive judgments show both anchoring and repulsion. *Psychological Science*, *33*(9), 1395-1407. https://doi.org/10.1177/09567976221089599

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599-637. https://doi.org/10.1111/cogs.12101

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105-136. https://doi.org/10.1016/j.cognition.2007.08.003

Zhu, J. Q., León-Villagrá, P., Chater, N., & Sanborn, A. N. (2022). Understanding the structure of cognitive noise. *PLoS Computational Biology*, *18*(8), e1010312. https://doi.org/10.1371/journal.pcbi.1010312

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments.*Psychological Review, 127*(5), 719–748. https://doi.org/10.1037/rev0000190

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times.*Psychological Review.* Advance online publication. https://doi.org/10.1037/rev0000427