COGNITIVE SCIENCE

Using large-scale experiments and machine learning to discover theories of human decision-making

Joshua C. Peterson[1]*, David D. Bourgin[1]†, Mayank Agrawal[2,3], Daniel Reichman[4], Thomas L. Griffiths[1,2]

Predicting and understanding how people make decisions has been a long-standing goal in many fields, with quantitative models of human decision-making informing research in both the social sciences and engineering. We show how progress toward this goal can be accelerated by using large datasets to power machine-learning algorithms that are constrained to produce interpretable psychological theories. Conducting the largest experiment on risky choice to date and analyzing the results using gradient-based optimization of differentiable decision theories implemented through artificial neural networks, we were able to recapitulate historical discoveries, establish that there is room to improve on existing theories, and discover a new, more accurate model of human decision-making in a form that preserves the insights from centuries of research.

Understanding how people make decisions is a central problem in psychology and economics (*1*–*3*). Having quantitative models that can predict these decisions has become increasingly important as automated systems interact more closely with people (*4*, *5*). The search for such models goes back almost 300 years (*6*) but intensified in the latter half of the 20th century (*7*, *8*) as empirical findings revealed the limitations of the idea that people make decisions by maximizing expected utility (EU) (*9*–*11*). This led to the development of new models such as prospect theory (PT) (*8*, *12*). Recently, this theory-driven enterprise has been complemented by data-driven research using machine learning to predict human decisions (*13*–*19*). Although machine learning has the potential to accelerate the discovery of predictive models of human judgments (*20*–*22*), the resulting models are limited by small datasets and are often uninterpretable (*5*). To overcome these challenges, we introduce a new approach based on defining classes of machine-learning models that embody constraints based on psychological theory. We present the largest experiment studying people's choices to date, allowing us to use our approach to systematically evaluate existing theories, identify a lower bound on optimal prediction performance, and propose a new descriptive theory that reaches this bound and contains classic theories as special cases.

We focus on risky choice, one of the most basic and extensively studied problems in decision theory (*8*, *23*). Risky choice has largely been examined using "choice problems," scenarios in which decision-makers face a choice between two gambles, each of which has a set of outcomes that differ in their payoffs and probabilities (Fig. 1A). Researchers studying risky choice seek a theory, which we formalize as a function that maps from a pair of gambles, A and B, to the probability $P(A)$ that a decision-maker chooses gamble A over gamble B, that is consistent with human decisions for as many choice problems as possible. Discovering the best theory is a formidable challenge for two reasons. First, the space of choice problems is large. The value and probability of each outcome for each gamble define the dimensions of this space, meaning that describing a pair of gambles could potentially require dozens of dimensions. Second, the space of possible theories is even larger, with theories of choice between two options spanning all possible functions mapping choice problems in $\mathbb{R}^{2d}$ to $\mathbb{R}$, i.e., from a vector of d gamble outcomes and d associated probabilities to a choice probability.

Machine-learning methods such as deep neural networks (*24*) excel at function approximation (*25*, *26*) and thus provide a tool that could potentially be used to automate theory search. However, these methods typically require large amounts of data. Historically, datasets on risky choice have been small: Influential papers focused on a few dozen choice problems (*27*) and the largest previous dataset featured <300 (*28*). Consequently, off-the-shelf methods have performed poorly in predicting human choices (*29*). Furthermore, even when data are abundant, the functions discovered by machine-learning algorithms are notoriously hard to interpret (*30*), making for poor explanatory scientific models.

To address these challenges, we collected a large dataset of human decisions for almost 10,000 choice problems presented in a format that has been used in previous evaluations of models of decision-making (*27*–*29*) (Fig. 1A). This dataset includes >30 times the number of problems in the largest previous dataset (*27*) (Fig. 1B). We then used this dataset to evaluate differentiable decision theories that exploit the flexibility of deep neural networks but use psychologically meaningful constraints to pick out a smooth, searchable landscape of candidate theories with shared assumptions. Differentiable decision theories allow the intuitions of theorists to be combined with gradient-based optimization methods from machine learning to broadly search the space of theories in a way that yields interpretable scientific explanations.

More formally, we define a hierarchy over decision theories (Fig. 1C) reflecting the addition of an increasing number of constraints on the space of functions. These constraints express psychologically meaningful theoretical commitments. For example, one class of theories contains all functions in which the value that people assign to one gamble can be influenced by the contents of the other gamble. If theories in this class are more predictive than those that belong to the simpler classes contained within it (e.g., where the value of gambles are independent), then we know that these simpler theories should be eliminated. We enforce each constraint by modifying the architecture of artificial neural networks, resulting in differentiable decision theories. This theory-driven approach to defining constraints contrasts with generic methods for constraining neural networks, such as restricting their size or the ranges of their weights (*31*). After optimizing a differentiable theory to best fit human behavior, it will ideally have picked out the optimal theory in its class.

The lowest levels of our hierarchy contain the simplest theories, including classic models of choice. Objectively, gambles that yield higher payouts in the long run are those with higher expected value (EV), with the value $V(A)$ of gamble A being $\sum_i x_i p_i$, where outcome i of gamble A has payoff x_i and probability p_i. In our hierarchical partitioning, this is the simplest possible theory because it has no descendants. Moving up the hierarchy, and following expected utility (EU) theory (*6*, *32*), we can ask the question of whether payouts x_i are viewed subjectively by decision-makers: $V(A) = \sum_i u(x_i) p_i$. When $u(\bullet)$ is the identity function $u(x) = x$, EU reduces to EV and thus contains it. Theories based on EU have historically relied on explicit proposals for the form of $u(\bullet)$, which are typically simple, nonlinear parametric functions (*33*). By contrast, we search the entire class by learning the optimal $u(\bullet)$ with a neural network (we call the resulting model "neural EU"), and use automatic differentiation to optimize the model $P(A) \propto \exp\{\eta \sum_i u(x_i) p_i\}$, where η captures the degree of determinism in people's responses (*34*). This can be viewed as a neural network

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA. [2]Department of Psychology, Princeton University, Princeton, NJ 08540, USA. [3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA. [4]Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609, USA.
*Corresponding author. Email: joshuacp@princeton.edu
†Present affiliation: Spotify.

Corrected 11 June 2021. See full text.

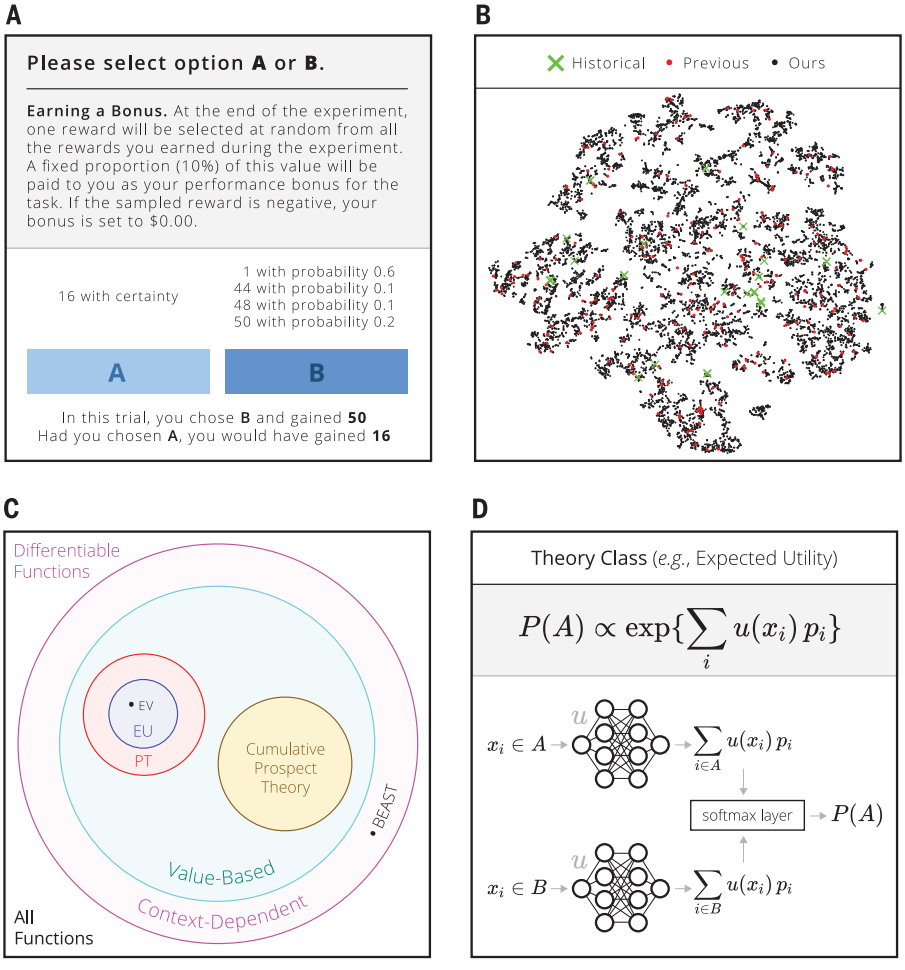

Fig. 1. Applying large-scale experimentation and theory-driven machine learning to risky choice. (**A**) Experiment interface in which participants made choices between pairs of gambles ("choice problems") and were paid at the end of the experiment based on their choice in a single randomly selected gamble. (**B**) Each pair of gambles can be described by a vector of payoffs and probabilities. Reducing the resulting space to two dimensions (2D) allows us to visualize coverage by different experiments. Each point is a different choice problem, and colors show reconstructions of the problems used in influential experiments (green), the previous largest dataset (red), and our 9831 problems, which provide much broader coverage of the problem space. This 2D embedding results from applying *t*-distributed stochastic neighbor embedding (*t*-SNE) to the hidden-layer representation of our best-performing neural network model. (**C**) We define a hierarchy of theoretical assumptions expressed as partitions over function space that can be searched. More complex classes of functions contain simpler classes as special cases, allowing us to systematically search the space of theories and identify the impact of constraints that correspond to psychologically meaningful theoretical commitments. All model classes are described in the main text. (**D**) Differentiable decision theories use the formal structure of classic theories to constrain the architecture of the neural network. For example, our EU model uses a neural network to define the utility function but combines those utilities in a classic form, resulting in a fully differentiable model that can be optimized by gradient descent.

architecture in which the output layer is a softmax function $e^{\eta z_j}/\Sigma_k e^{\eta z_k}$, there is one node for each gamble, the hidden units in the second-to-last layer encode the utilities of the outcomes, and the final layer of weights corresponds to their probabilities (Fig. 1D). Figure 2 shows that the discovered form of $u(\bullet)$ is similar to those proposed by human theorists (i.e., decreasing marginal utility and asymmetry) but outperforms any of those theories and can be learned using only a quarter of our data. [All theories are evaluated on their cross-validated generalization performance, meaning that model complexity is already implicitly accounted for in our analyses; we focus on mean-squared error (MSE) for consistency with previous evaluations of models of decision-making (*28*, *29*) but also include analyses of cross-entropy in the supplementary materials.] The decision preference accuracy [i.e., the proportion of problems in which the model prediction for $P(A)$ is >0.5 when the observed proportions are also >0.5] for this model was 81.41%.

Next, mirroring subjective EU (*7*) and PT, we can ask the question of whether the probabilities (p_i) are also viewed subjectively by decision-makers: $v(A) = \Sigma_i u(x_i)\pi(p_i)$. Again, $\pi(\bullet)$ can take on classic forms or be learned from data ("neural PT"). Figure 2B shows that a form of $\pi(\bullet)$ that outperforms all proposals by human theorists can be learned using one-fifth of our data and exhibits overweighting of events with medium to low probability. This overweighting is much smaller than is typically found in applications of PT, in part reflecting the difference in the range of choice problems that we consider relative to classic studies. We will return to this point later. The decision preference accuracy for this model was 82.33%.

Allowing separate $\pi(\bullet)$ functions for positive and negative outcomes, respectively, and applying them cumulatively to an ordered set of outcomes corresponds to the most popular modern variant of PT: cumulative PT (CPT) (*12*, *35*) (Fig. 2B; see the materials and methods). Notably, "neural CPT" does not contain neural PT because the former cannot violate stochastic dominance. With small amounts of data, corresponding to the largest previous experiments (*28*, *29*), CPT outperforms PT, accounting for its popularity. However, this trend reverses as the amount of data is increased, which illustrates that suitably large datasets, in addition to aiding machine learning, provide more robust evaluation.

Next, we can ask whether the possible outcomes of a gamble affect the perception of each other and their probabilities and vice versa. More formally, we learn a neural network $f(\bullet,\bullet)$ such that $P(A) \propto \exp\{f(x_A,p_A)\}$, where x_A and p_A are the vector of payoffs and probabilities associated with gamble A, respectively. This function class computes the value of a gamble ("value-based") like PT and others but does not enforce linearity when combining payoffs and probabilities. Notably, this class of models includes those that violate the independence axiom in decision-making (*32*). Figure 3A shows that there exists a value-based theory that results in a greater improvement in performance over PT than PT does over EU.

Relaxing the constraint that each gamble is valued independently results in our most general class of functions, "context-dependent" functions $g(\bullet)$ where $P(A) = g(x_A, p_A, x_B, p_B)$. This class of models includes those that violate both the independence and transitivity axioms in decision-making (*32*). This formulation provides a way to estimate the performance of the

Corrected 11 June 2021. See full text.

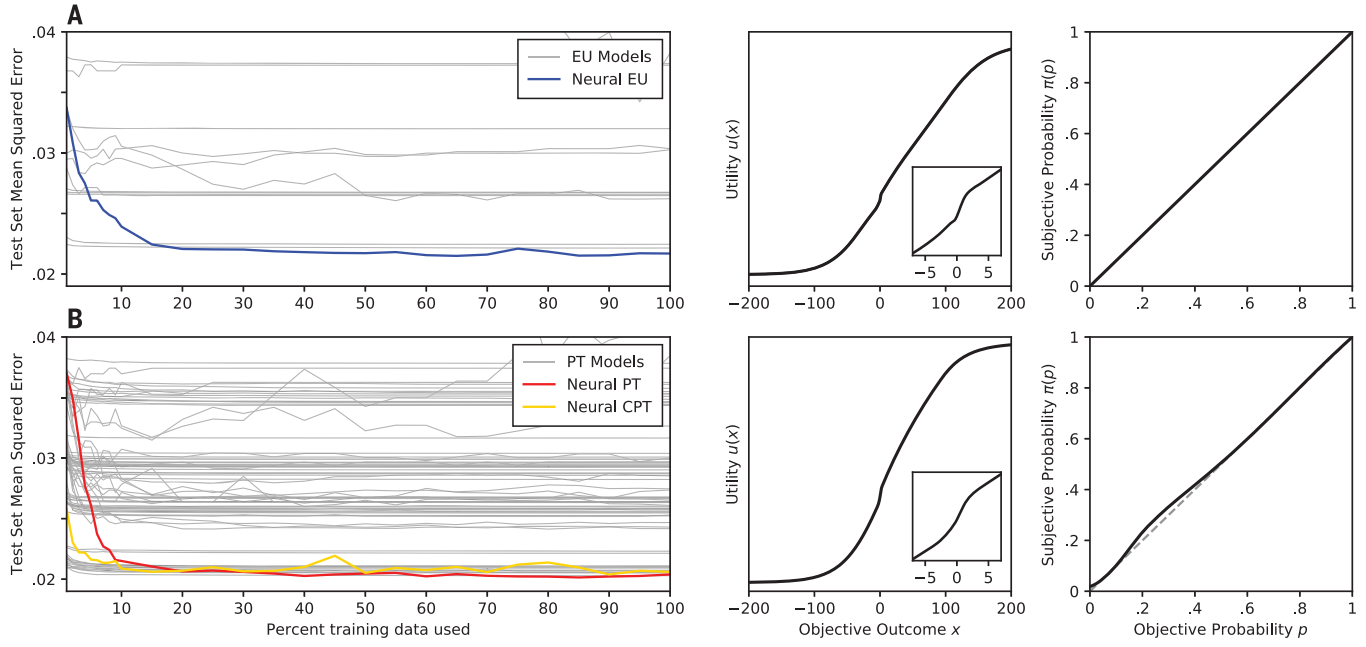

Fig. 2. Comparing classic theories proposed by human researchers with differentiable decision theories discovered through machine learning. (**A**) EU. In the left panel, EU with a learned optimal utility function outperforms classic models (gray lines) given enough data. Performance is assessed in terms of prediction error (MSE) on ~1000 unseen choice problems as a function of the amount of training data used for model fitting (~9000 choice problems, average of 50 runs). The right panel shows that the utility function identified through this optimization method reproduces many of the characteristics suggested by human theorists. (**B**) PT. In the left panel, PT with learned utility and probability weighting functions outperforms classic models (gray lines) given enough data. CPT, the modern variant of PT, performs better with small amounts of data, on the scale of previous experiments, but slightly worse with more data. The right panel shows the optimized utility and probability weighting functions for PT.

near-optimal theory because it has no constraints (except that it must be differentiable). In particular, this model is simply a fully unconstrained neural network that takes all information about both gambles as input and outputs $P(A)$. The move from value-based to context-dependent functions results in the largest improvement yet in prediction performance and a decision preference accuracy of 84.81%. Because value-based theories assign values to gambles independently, comparing their performance with this less-constrained neural network implicitly asks whether the value of gambles is contextual (i.e., whether one gamble and its parameters affect one's perception of the other gamble); our results indicate that this is the case and, further, that the impact of this effect is substantial.

One context-dependent psychological model that has outperformed others in recent evaluations is a process model called Best Estimate and Sampling Tools (BEAST) (*27*), which makes decisions by combining EV with the results of four different kinds of simulations, a subset of which integrates information about both outcomes and probabilities across gambles (*27*). Despite outperforming all other models with small amounts of data, Fig. 3A shows that BEAST falls short of our best-performing model in the context-dependent class as the amount of data increases. Further, we found that several other models from diverse modeling traditions in risky choice that also incorporate various forms of context performed no better than PT on our dataset (see Fig. 3B and the supplementary materials). We also replicated our core findings in a second dataset of 1000 problems collected from a new group of participants, including showing that we see similar results when evaluating the models at the level of individual participants (see the supplementary materials).

Having evaluated the most competitive theories at each level of our hierarchy, we find that the best-performing theory belongs to the most complex class we defined and lies outside of all simpler classes. This implies that the best predictions of human choices result from viewing payoffs and their probabilities subjectively but, more importantly, in ways that are sensitive to the context of the competing gamble. However, as a relatively unconstrained neural network, this model provides limited psychological insight and is highly susceptible to overfitting the noise in small datasets because of its expressive power.

To better understand which aspects of context were responsible for better model performance, we conduct a second pass of our method. In particular, we define a class of models ("contextual multiplicative") where $V(A) = \Sigma_{i \in A} u(x_i, c_1)\pi(p_i, c_2)$ and c_1 and c_2 are vectors potentially containing information from x_A, x_B, p_A, and p_B that condition $u(\bullet)$ and $\pi(\bullet)$, allowing subjective rewards and probabilities to vary depending on the problem (see the supplementary materials for more details). Results are shown in Fig. 3A. When conditioning utilities on other outcomes within a gamble ($c_1 = x_A$; "intra-gamble outcome context"), performance improves only marginally and does not match the value-based class. However, when instead conditioning utilities on all other outcomes across both gambles ($c_1 = \{x_A, x_B\}$; "inter-gamble outcome context"), performance improves markedly. Further allowing probabilities to interact ($c_2 = \{p_A, p_B\}$; "inter-gamble outcome/probability context") provides marginal improvement. Existing theories that fall into these classes performed no better than PT (i.e., where c_1 and c_2 contain no information; see the supplementary materials). Finally, a "fully contextual multiplicative" model, where $c_1 = c_2 = \{x_A, x_B, p_A, p_B\}$ fails to improve performance, likely because it has twice as many parameters as the context-dependent model as a result of conditioning two networks on all information from both gambles. This analysis helps to illuminate the key properties of the best-performing theory found within the context-dependent class: Outcomes and probabilities are largely combined

Corrected 11 June 2021. See full text.

multiplicatively to form an average but are subjectively transformed in ways that depend on information across both gambles, especially outcomes.

To identify a theory that has these characteristics, and is more continuous with previous accounts of human decision-making, and able to be trained with fewer data, we compared EU theory with the best-performing context-dependent model, identifying patterns in the types of problems where PT performed less well. We found that the largest differences concerned dominated gambles (i.e., where all outcomes of one gamble are better than all outcomes of the other) and pairs of gambles that pit likely losses against high uncertainty (see the supplementary materials). To address this, we defined a new model based on the hypothesis that people use different strategies for different choice problems, reminiscent of the dual process theories of risky choice (*36*). These strategies can correspond to classic models, preserving their insights. Further, if we allow the selection of these strategies based on the properties of both gambles, then this model is a simpler, more specific form of the contextual multiplicative class, which normally selects from a theoretically infinite set of models (i.e., utility and weighting functions) instead of a relatively compact, fixed set.

The resulting mixture of theories (MOT) model learns to apply one of two utility functions and one of two probability weighting functions. Two accompanying neural networks take both gambles as input and output convex mixture weights for both subjective functions, which are learned jointly in a "mixture of experts" architecture (*37*). For dominated gambles, the gamble values determined by these mixtures are bypassed and a learned, fixed probability of choosing the dominated gamble is taken as the prediction. Figure 4A shows that this model generalizes to unseen problems as well as our best, fully unconstrained neural network (context-dependent), with a similar decision preference accuracy of 84.15%, and is able to achieve better performance with fewer data. The functional forms rediscovered by the mixture model correspond closely to classic EU and PT models (Fig. 4B). In particular, one learned utility function clearly produces loss aversion (*8*) and one probability weighting function overweights low probabilities in a way that is more consistent with the classic effects modeled by PT (*8*). By looking at the problems where the different components are used, we can see when these characteristics are most important to capturing human decisions (Fig. 4, lower panels). The best predictors of the utility function used, which an ablation analysis reveals is the most important context effect (see the supplementary materials), were maximum outcome, minimum outcome, and outcome variability. The best predictors for probability weighting functions were minimum outcome and number of losses (see the supplementary materials). This focus on the context of outcomes across gambles is corroborated by similar performance being produced by a variant of MOT in which the mixture network is only given outcomes as input (Fig. 4A, dashed green line). The clusters of problems assigned to each subjective function emerge in a representation based on hidden activations from the context-dependent model, suggesting that the MOT is capturing a structure similar to this unconstrained model.

We also found that MOT provides a competitive model of individual-level behavior. In particular, we fine-tuned the parameters of the MOT model above trained on aggregate data to the behavior of individual participants, obtaining an average MSE of 0.058 on out-of-sample decisions for each participant. For comparison, we repeated this procedure using the best-performing parametric form of PT, a common choice for individual-level modeling of risky choice, obtaining an average MSE of 0.063. See the supplementary materials for additional details and results.

Our results illustrate the successes of human ingenuity, in particular, finding good functional forms for the EU and PT models. However, they also illustrate that this ingenuity can be supplemented by an automated search over models given enough data, and that as the class of models becomes less restrictive and the dataset becomes larger, this automated approach begins to substantially outperform the best models of decision-making developed by human researchers. This does not mean that theories developed by psychologists and economists are

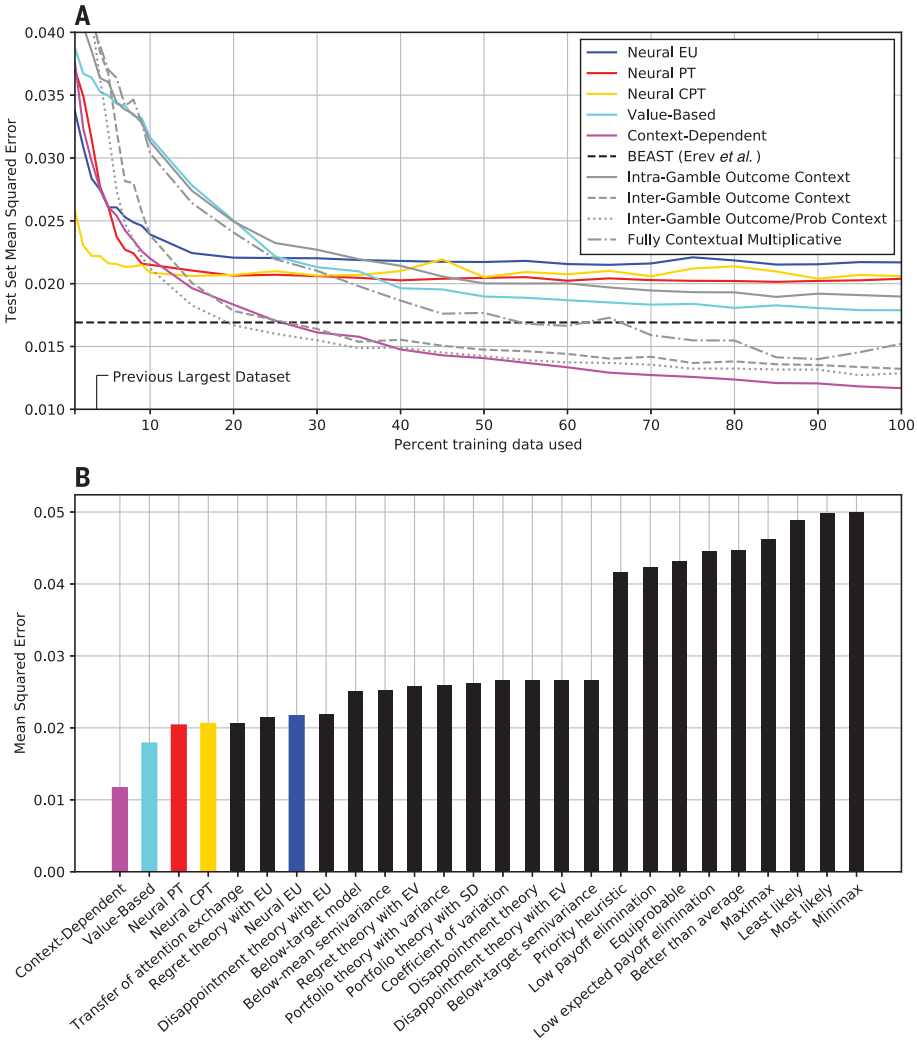

Fig. 3. Complex decision theories exhibit better predictive performance than simpler ones. (**A**) Performance of differentiable decision theories. As model flexibility increases, performance increases, along with data requirements. (**B**) Performance comparison between differentiable decision theories and 21 other well-known theories across the risky choice literature, none of which outperforms PT (see the supplementary materials for more details).

Corrected 11 June 2021. See full text.

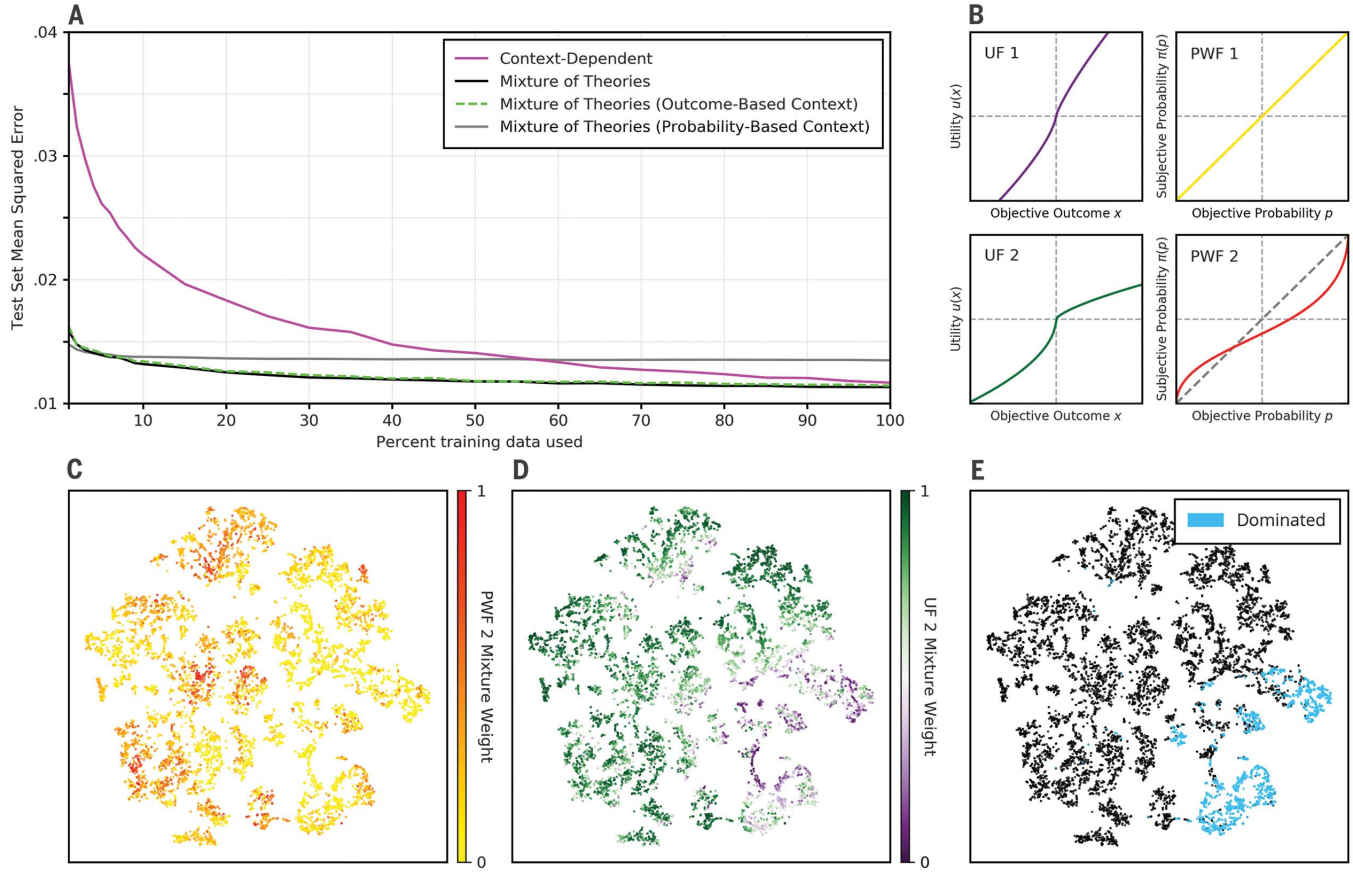

Fig. 4. The MOT model assigns a fixed learned decision probability for gambles that are dominated and determines the value of gambles in all other choice problems using mixtures of two utility and two probability-weighting functions. (**A**) MOT obtains slightly better performance than the best learned decision theory, even when the mixture networks and thus context effects are limited to information about outcomes only. (**B**) Components of the model are interpretable by design and resemble the classic forms for EU and PT: Utility function 1 (UF 1) is symmetric, whereas UF 2 shows loss aversion. Probability weighting function 1 (PWF 1) is linear, whereas PWF 2 shows overweighting of lower probabilities. 2D *t*-SNE embeddings of the choice problems based on the hidden unit representations of the best-performing context-dependent neural network exhibit distinct clusters that correspond to problems where MOT applied UF 2 (**C**), PWF 2 (**D**), and the dominated component (**E**).

not valuable for predicting human behavior, even in the big-data regime: Our MOT model was able to leverage large amounts of data to learn which classic theory best describes human behavior in different contexts. These models, variants of EU and PT, have been subject to rigorous mathematical analyses over decades, which are preserved in this new theory.

Further, whereas MOT is a special case of the contextual multiplicative model, it performs better due to theory-driven simplifications and is thus more sample efficient. Human ingenuity will also be required for potentially translating this descriptive theory into normative and process models (*38*, *39*). We anticipate that our approach of defining differentiable theories that express meaningful psychological constraints can be applied in other settings as we continue to gather more data on human decision-making.

Finally, it is noteworthy that models of decision-making developed by human researchers tend to outperform our machine-learning models when we only consider amounts of data that are consistent with the scale of previous behavioral research, but this trend reverses when more data are available. This pattern may imply that the complexity of psychological theories has been constrained by limited data. As we begin to move into a regime of big behavioral data, our theories are going to have to become increasingly complex to be able to capture the systematic variation that these larger datasets reveal. The use of large datasets (*40*, *41*) has revolutionized machine learning, computer vision, and artificial intelligence. Our study is one of the first to use a similar methodology in systematically investigating theories of human cognition. We believe that the use of large datasets coupled with machine-learning algorithms offers enormous potential for uncovering new cognitive and behavioral phenomena that would be difficult to identify without such tools (*42*).

REFERENCES AND NOTES

1. N. C. Barberis, *J. Econ. Perspect.* **27**, 173–196 (2013).
2. R. Hastie, R. M. Dawes, *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making* (Sage, 2009).
3. I. Gilboa, *Theory of Decision Under Uncertainty* (Cambridge Univ. Press, 2009), vol. 45.
4. A. Jameson, "Choices and decisions of computer users," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. A. Jacko, Ed. (CRC Press, 2012), pp. 77–94.
5. I. Rahwan *et al.*, *Nature* **568**, 477–486 (2019).
6. D. Bernoulli, *Econometrica* **22**, 23–36 (1954).
7. L. J. Savage, *The Foundations of Statistics* (Courier, 1972).
8. D. Kahneman, A. Tversky, *Econometrica* **47**, 263–292 (1979).
9. M. Allais, *Econometrica* **21**, 503–546 (1953).
10. H. J. Einhorn, R. M. Hogarth, *J. Bus.* **59** (S4), S225–S250 (1986).
11. D. Ellsberg, *Q. J. Econ.* **75**, 643–669 (1961).
12. A. Tversky, D. Kahneman, *J. Risk Uncertain.* **5**, 297–323 (1992).
13. D. Fudenberg, J. Kleinberg, A. Liang, S. Mullainathan, Measuring the completeness of theories. arXiv:1910.07022 [econ.TH] (15 October 2019).
14. G. Noti, E. Levi, Y. Kolumbus, A. Daniely, Behavior-based machine-learning: A hybrid approach for predicting human decision making. arXiv:1611.10228 [cs.LG] (30 November 2016).

15. T. Yarkoni, J. Westfall, *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
16. A. Peysakhovich, J. Naecker, *J. Econ. Behav. Organ.* **133**, 373–384 (2017).
17. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
18. J. S. Hartford, J. R. Wright, K. Leyton-Brown, *Adv. Neural Inf. Process. Syst.* **29**, 2424–2432 (2016).
19. L. He, P. P. Pantelis, S. Bhatia, *Manage. Sci.*, in press.
20. J.-F. Bonnefon, A. Shariff, I. Rahwan, *Science* **352**, 1573–1576 (2016).
21. A. Rosenfeld, S. Kraus, "Predicting human decision-making: From prediction to action," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, R. Brachman, F. Rossi, P. Stone, Eds. (Morgan & Claypool, 2018), vol. 12, no. 1, pp. 1–150; https://doi.org/10.2200/S00820ED1V01Y201712AIM036.
22. C. F. Camerer, "Artificial intelligence and behavioral economics," in *The Economics of Artificial Intelligence: An Agenda*, A. Agrawal, J. Gans, A. Goldfarb, Eds. (Univ. of Chicago Press, 2018), pp. 587–608.
23. W. Edwards, *Psychol. Bull.* **51**, 380–417 (1954).
24. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436–444 (2015).
25. G. Cybenko, *Math. Contr. Signals Syst.* **2**, 303–314 (1989).
26. K. Hornik, *Neural Netw.* **4**, 251–257 (1991).
27. I. Erev, E. Ert, O. Plonsky, D. Cohen, O. Cohen, *Psychol. Rev.* **124**, 369–409 (2017).
28. O. Plonsky *et al*., Predicting human decisions with behavioral theories and machine learning. arXiv:1904.06866 [cs.AI] (15 April 2019).
29. O. Plonsky, I. Erev, T. Hazan, M. Tennenholtz, "Psychological forest: Predicting human behavior," in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, CA, 4–9 February 2017; https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14925.
30. F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 [stat.ML] (28 February 2017).
31. J. Schmidhuber, *Neural Netw.* **61**, 85–117 (2015).
32. J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, 1944).
33. P. P. Wakker, *Prospect Theory: For Risk and Ambiguity* (Cambridge Univ. Press, 2010).
34. D. McFadden, "Conditional logit analysis of qualitative choice behaviour," in *Frontiers in Econometrics*, P. Zarembka, Ed. (Academic, 1973), pp. 105–142.
35. H. Fennema, P. Wakker, *J. Behav. Decis. Making* **10**, 53–64 (1997).
36. K. Mukherjee, *Psychol. Rev.* **117**, 243–255 (2010).
37. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, *Neural Comput.* **3**, 79–87 (1991).
38. N. Stewart, N. Chater, G. D. Brown, *Cognit. Psychol.* **53**, 1–26 (2006).
39. R. Bhui, S. J. Gershman, *Psychol. Rev.* **125**, 985–1001 (2018).
40. J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
41. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
42. T. L. Griffiths, *Cognition* **135**, 21–23 (2015).

ACKNOWLEDGMENTS

We thank F. Callaway, N. Daw, P. Ortoleva, R. Dubey, and reviewers for providing comments on the manuscript. Preliminary analyses of our dataset were presented at the International Conference on Machine Learning. The analyses presented here are entirely new and go substantially beyond that early work. **Funding:** This work was supported by the Future of Life Institute, the Open Philanthropy Foundation, the NOMIS Foundation, DARPA (cooperative agreement D17AC00004), and the National Science Foundation (grant no. 1718550). **Author contributions:** Conceptualization: J.C.P., T.L.G., D.D.B., D.R.; Data curation: D.D.B.; Formal analysis: J.C.P., D.D.B., M.A., T.L.G.; Funding acquisition: T.L.G.; Investigation: J.C.P., D.D.B., M.A.; Methodology: J.C.P., T.L.G.; Project administration: D.R., T.L.G.; Software: J.C.P., D.D.B.; Supervision: T.L.G., D.R.; Visualization: J.C.P., M.A.; Writing – original draft: J.C.P., M.A.; Writing – review and editing: J.C.P., T.L.G., D.D.B., M.A., D.R. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available to the public without registration at https://github.com/jcpeterson/choices13k.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/372/6547/1209/suppl/DC1
Materials and Methods
Figs. S1 to S11
Tables S1 to S7
References (*43*–*68*)
MDAR Reproducibility Checklist

10 August 2020; accepted 6 May 2021
10.1126/science.abe2629

Using large-scale experiments and machine learning to discover theories of human decision-making

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman and Thomas L. Griffiths

Science **372** (6547), 1209-1214.
DOI: 10.1126/science.abe2629originally published online June 10, 2021

Discovering better theories

Theories of human decision-making have proliferated in recent years. However, these theories are often difficult to distinguish from each other and offer limited improvement in accounting for patterns in decision-making over earlier theories. Peterson *et al.* leverage machine learning to evaluate classical decision theories, increase their predictive power, and generate new theories of decision-making (see the Perspective by Bhatia and He). This method has implications for theory generation in other domains.

Science, abe2629, this issue p. 1209; see also abi7668, p. 1150