science.sciencemag.org/content/372/6547/1209/suppl/DC1

Supplementary Materials for

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

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

*Corresponding author. E-mail: joshuacp@princeton.edu

Published 11 June 2021, *Science* **372**, 1209 (2021)
DOI: 10.1126/science.abe2629

This PDF file includes:

Materials and Methods
Figs. S1 to S11
Tables S1 to S7
References

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/372/6547/1209/suppl/DC1)

MDAR Reproducibility Checklist (PDF)

Materials and Methods

Choice problem generation

The `choices13k` dataset contains 13,006 risky choice problems. Each problem was uniquely identified by a set of 11 problem features, including the parameters of the payoff distributions for the two gambles (rewards could be sampled from either uniform or skewed distributions), whether or not the participant was shown the reward probabilities associated with each gamble before making their decision, the reward correlation between the two gambles (positive, negative, or zero), and whether or not the participant received feedback on the reward they received or missed out on after making a selection. Values for each feature were sampled following the process used in (*27*): for example, the expected value for the first gamble was sampled from a Uniform distribution on the interval from $[10, 30]$, the shape of the payoff distribution in Gamble B was sampled from a categorical distribution with a 25% chance of symmetric, 50% chance of degenerate, and 25% chance of left or right-skew, etc. For complete details, see (*27*), Appendix D. All of our main analyses ignored "ambiguous" problems (i.e., problems where the gamble reward probabilities were masked when presented to participants), as well as those without feedback, resulting in a dataset of 9,831 problems.

Data collection

Human selection frequencies on the problems in the `choices13k` dataset were collected using Amazon Mechanical Turk. Informed consent was obtained from all individual participants included in the study, which was IRB-approved (CPHS Protocol number 2015-05-7551: "Cognitive Research Using Amazon Mechanical Turk"). Each participant made selections on 20 choice problems in blocks of five trials per problem. Participants were paid a base rate of \$0.75 USD for their participation plus a bonus proportional to the reward from a single gamble sampled from their history at the end of the experiment. The gamble selection interface was made to match the interface used in previous choice prediction competitions (*27*) as closely as possible.

On each trial, participants were asked to select the most appealing option from a set of two gambles. Both gambles were represented to participants as a list of rewards and their associated outcome probabilities. One gamble (henceforth "Gamble A") was constrained to have only two outcomes. The second gamble ("Gamble B") yielded a fixed reward with probability p and the outcome of a lottery (i.e., the outcome of another explicitly described gamble) otherwise. The probabilities and rewards associated with each gamble were dictated according to the values of the parameters for that problem. Upon selection of a gamble, a reward was sampled from that gamble's payoff distribution and added to the subject's reward history (used to determine bonuses). In the feedback condition participants were told the value of the rewards they received and missed out on after each selection, while in the no-feedback condition this information was suppressed.

The presentation format for the gambling problems closely followed that of the 2015 and 2018 Choice Prediction Competitions (*27, 28*), with the exception that the block parameter was reduced from five to two, and feedback and no-feedback blocks were not required to be presented sequentially. For a detailed description of the gamble presentation and problem parameterization, see (*27*), Appendices A & C.

Data Filtering and Behavioral Analyses

Participants were required to be from the US and have completed at least 500 tasks with a 95% acceptance rate or higher on the platform. We then dropped any participants that had over 80% of their selections as the left/right gamble, resulting in a total of $14,711$ participants. This filtering was completed before any analysis or modeling of the data.

In the filtered dataset, majority preferences in aggregate choice rates violated stochastic dominance in 14% of problems. Participants chose the gamble with a higher expected value for 77.6% of problems.

Modeling problem

The modeling goal using the data described above is to predict the probability $P(A)$ that human decision makers will choose gamble A over gamble B, where $P(B) = 1 - P(A)$. It should be noted that this problem is more difficult than finding a model that produces categorical decision preferences that match the majority of human responses, since a model could output *e.g.,* $P(A) = 0.51$, and still be in agreement with the overall preference of humans where say $P(A) = 0.99$. Instead, we want a model that matches the human proportions as closely as possible, since those proportions represent additional information, potentially including: (1) the relative psychological values of each gamble, (2) the degree to which decision makers respond deterministically, (3) the propensity for decision makers to disagree for a particular problem.

To measure error between model outputs $\hat{P}(A)$ and human response proportions $P(A)$, we use mean squared error (MSE), following past work (*28, 29*):

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{n}[\hat{P}_n(A) - P_n(A)]^2 \tag{1}$$

where n ranges over problems and N is the total number of problems. However, since we are comparing probability distributions, it may be more appropriate to measure error via cross entropy:

$$\mathcal{L}_{CP} = -\sum_{n} P_n(A) \cdot \log \hat{P}_n(A) + P_n(B) \cdot \log \hat{P}_n(B). \tag{2}$$

Since both $\mathcal{L}_{MSE}$ and $\mathcal{L}_{CP}$ error functions produced a similar pattern of results and $\mathcal{L}_{MSE}$ has been emphasized in previous work, we report results for $\mathcal{L}_{MSE}$ in the main text and present a subset of those for $\mathcal{L}_{CP}$ in Supplementary Fig. S5 below.

Model formulations

All models evaluated in this paper except heuristics and the "Context-Dependent", both explained below, output values $V(A)$ and $V(B)$ for gambles A and B respectively. The probability that human decision makers will choose gamble A over gamble B is then defined as

$$P(A) = \frac{e^{\eta V(A)}}{e^{\eta V(A)} + e^{\eta V(B)}} \; , \tag{3}$$

where η is a free parameter that controls the level of deterministic responding in addition to the ratio of psychological values $V(A)/V(B)$. Note that the above definition implies that $P(A) \propto e^{\eta V(A)}$.

Expected Value models

The Expected Value (EV) of a gamble is the long-run average outcome value, defined as $V(A) = \sum_i x_i \, p_i$, and thus its corresponding model contains no free parameters other than η, which does not affect the values of each gamble.

Expected Utility models

Expected Utility (EU) models (*6, 32*) can incorporate arbitrary transformations $u(\cdot)$ of outcomes x, such that their utility is regarded subjectively, and $V(A) = \sum_i u(x_i) \, p_i$. Proposals for the form of $u(\cdot)$ are typically simple, nonlinear monotonic parametric functions (*33*). We evaluated the following 12 fixed and parametric forms of $u(\cdot)$:

1. Linear: $u(x) = \lambda x$
2. Linear Loss-Averse: $u(x) = \begin{cases} x & \text{if } x \geq 0 \\ \lambda x & \text{if } x < 0 \end{cases}$

3. Power (*12, 43, 44*): $u(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-x)^{-\beta} & \text{if } x < 0 \end{cases}$

4. Exponential (*12, 45*): $u(x,\alpha) = \begin{cases} 1 - e^{-\alpha x} & \text{if } x \geq 0, \alpha > 0 \\ x & \text{if } x \geq 0, \alpha = 0 \\ e^{-\alpha x} - 1 & \text{if } x \geq 0, \alpha < 0 \\ -\lambda(1 - e^{-\alpha x}) & \text{if } x < 0, \alpha > 0 \\ -\lambda x & \text{if } x < 0, \alpha = 0 \\ -\lambda(e^{-\alpha x} - 1) & \text{if } x < 0, \alpha < 0 \end{cases}$

5. Normalized Exponential (*46*): $u(x) = \begin{cases} \frac{1}{\alpha}(1 - e^{-\alpha x}) & \text{if } x \geq 0 \\ \frac{-\lambda}{\beta}(1 - e^{\beta x}) & \text{if } x < 0 \end{cases}$

6. Normalized Logarithmic (*46*): $u(x) = \begin{cases} \frac{1}{\alpha}\log(1 + \alpha x) & \text{if } x \geq 0 \\ \frac{-\lambda}{\beta}\log(1 - \beta x) & \text{if } x < 0 \end{cases}$

7. Normalized Power (*46*) $u(x) = \begin{cases} ((1+\alpha)x)^{(1+\alpha)} & \text{if } x \geq 0 \\ -(-(1+\beta)\frac{x}{\lambda})^{(1+\beta)} & \text{if } x < 0 \end{cases}$

8. Quadratic: $u(x) = \begin{cases} \alpha x - x^2 & \text{if } x \geq 0 \\ \lambda(-\beta x - x^2) & \text{if } x < 0 \end{cases}$

9. Logarithmic: $u(x) = \begin{cases} \log(\alpha + x) & \text{if } x \geq 0 \\ -\lambda\log(-\beta x) & \text{if } x < 0 \end{cases}$

10. Exponential Power (*47, 48*): $u(x) = \begin{cases} \gamma - e^{-\beta x^{\alpha}}) & \text{if } x \geq 0 \\ -\lambda(\gamma - e^{-\beta \times (-x)^{\alpha}})) & \text{if } x < 0 \end{cases}$

11. General Linear (*12, 45*): $u(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ \lambda\beta x & \text{if } x < 0 \end{cases}$

12. General Power (*12, 45*): $u(x) = \begin{cases} \beta * x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-\delta x)^{\gamma} & \text{if } x < 0 \end{cases}$

Notably, every utility function $u(x)$ above except "Linear" is a piecewise function that allows for asymmetry about $x = 0$ depending on whether $x \geq 0$ or $x < 0$. In particular, when $x \geq 0$, a separate parameter may control the shape of the curve in the loss domain, and a non-negative parameter λ is always learned for the loss domains such that $u(x) = -\lambda u(-x)$ (*12, 33*). For each model, we estimate their possible parameters λ, α, β, γ, and δ, and as well as η for the overall resulting model.

Neural Expected Utility

To search the entire class of possible utility functions that might best characterize our data, we estimate η as well as the parameters of a neural network $u(x)$ for the overall model, which we call "Neural EU". In particular, our neural network comprised a single input unit, 10 hidden units, and a single output unit, all with sigmoid activation functions. Networks larger than this did not improve performance.

Prospect Theory models

Prospect Theory (PT) models (*8*) incorporate transformations $u(\cdot)$ of outcomes x as in EU, as well as transformations $\pi(\cdot)$ of probabilities p, such that utilities are probabilities are regarded subjectively, and $V(A) = \sum_i u(x_i)\,\pi(p_i)$. Proposals for the forms of $u(\cdot)$ and $\pi(\cdot)$ are typically simple, nonlinear monotonic parametric functions (*8, 12, 33, 49*). We evaluated the following 8 fixed and parametric forms of $\pi(\cdot)$:

1. Kahneman-Tversky (*12*): $\pi(p) = \frac{p^\alpha}{(p^\alpha + (1-p)^\alpha)^{\frac{1}{\alpha}}}$

2. Log Odds Linear (*50*): $\pi(p) = \frac{\beta p^\alpha}{\beta p^\alpha + (1-p)^\alpha}$

3. Power (*51*): $\pi(p) = \beta p^{\alpha}$

4. NeoAdditive (*33*): $\pi(p) = \beta + \alpha p$

5. Hyperbolic Log (*49, 52*): $\pi(p) = (1 - \alpha + \log p)^{\frac{\beta}{\alpha}}$

6. Exponential Power (*49, 52*): $\pi(p) = e^{-\frac{\alpha}{\beta}(1-p^{\beta})}$

7. Compound Invariance (*49, 53*): $\pi(p) = e^{\beta(-\log x)^{\alpha}}$

8. Constant Relative Sensitivity (*54*): $\pi(p) = \beta^{(1-\alpha)} + p^{\alpha}$

Each the above can be combined with all utility functions from the previous section. For each model, we estimate both the parameters of the utility functions and the possible parameters α and β for $\pi(\cdot)$, and as well as η for the overall resulting model.

Neural Prospect Theory

To search the entire class of possible $u(\cdot)$ and $\pi(\cdot)$ functions that might best characterize our data, we estimate η as well as the parameters of a neural networks $u(x)$ and $\pi(x)$ for the overall model, which we call "Neural PT". In particular, each neural network comprises a single input unit, 10 hidden units, and a single output unit, all with sigmoid activations functions. Networks larger than this did not improve performance.

Cumulative Prospect Theory models

An updated version of Prospect Theory, called Cumulative Prospect Theory (CPT), was proposed to additionally satisfy stochastic dominance (*12*), but also makes different predictions (*33*). In CPT, the value of any one gamble with outcomes x_i and probabilities p_i is given by

$$\sum_{i=1}^{k} \pi_i^- u\left(x_i\right) + \sum_{i=k+1}^{n} \pi_i^+ u\left(x_i\right). \tag{4}$$

where "decision weights" π_i^- and π_i^+ are applied in "cumulative" fashion to ordered outcomes $x_1 \leq \ldots \leq x_k \leq 0 \leq x_{k+1} \leq \ldots \leq x_n$:

$$\begin{aligned} \pi_1^- = \pi^-\left(p_1\right), \quad \pi_i^- = \pi^-\left(p_1 + \cdots + p_i\right) - \pi^-\left(p_1 + \cdots + p_{i-1}\right), \quad 2 \leq i \leq k, \\ \pi_n^+ = \pi^+\left(p_n\right), \quad \pi_i^+ = \pi^+\left(p_i + \cdots + p_n\right) - \pi^+\left(p_{i+1} + \cdots + p_n\right), \quad k+1 \leq i \leq n-1. \end{aligned} \tag{5}$$

In the neural variant of CPT ("Neural CPT"), as above, $\pi^-(\cdot)$, $\pi^+(\cdot)$, and $u(\cdot)$ are estimated via separate neural networks with single input/output units and 10 hidden units. Networks larger than this did not improve performance. The estimated utility function and probability weighting functions are illustrated in Fig. S10 and Fig. S11, respectively.

Value-Based models

A more flexible class of models are what we call Value-Based models, where $V(A) = f(\mathbf{x}_A, \mathbf{p}_A)$, where $\mathbf{x}_A$ and $\mathbf{p}_A$ are the vector of payoffs and probabilities associated with Gamble A respectively. We estimate $f(\cdot,\cdot)$ using a neural network with 64 hidden units with sigmoid activation functions and a single output unit. Networks larger than this did not improve performance. Unlike the above neural networks, which output utilities or decision weights, this network outputs $V(A)$ directly. Since the neural network integrates all information about a gamble at once, it can integrate outcomes and probabilities in any way it chooses, including in ways where the entire set of probabilities affect how each is weighed, the entire set of outcomes affects how each is weighted, and how weighting of either are affected by the other.

Context-Dependent models

Our most flexible class of models are neural networks that directly output $P(A)$ given all information about both gambles as input, in our case using two 32-unit hidden layers. Specifically, we define a neural network g such that

$$P(A) = g(\mathbf{x}_A, \mathbf{p}_A, \mathbf{x}_B, \mathbf{p}_B). \tag{6}$$

This structure allows for cases where $V(A)$, which is only implicitly represented in the network, can depend on the content of gamble B, and vice versa. For this reason, we call its corresponding model class “Context-Dependent”.

Contextual Multiplicative Models

One more restricted class of context-dependent models require that utilities and weighted probabilities within a gamble be combined multiplicatively (as in PT), yet allow subjective transformations of outcomes and probabilities to be dependent on context. This model maintains the interpretation of a gamble’s value as a subjectively weighted average of subjective rewards (again, as in PT), but allows subjective transformations to vary across different problems. More specifically, this model can be described generally by:

$$V(A) = \sum_{i \in A} u(x_i, \mathbf{c}_1)\, \pi(p_i, \mathbf{c}_2), \tag{7}$$

where $\mathbf{c}_1$ is a context vector that conditions the utility function $u(\cdot)$, and $\mathbf{c}_2$ is a context vector that conditions the probability weighting function $\pi(\cdot)$. We implement such models using neural network utility and weighting functions that additionally take a context vector as input (concatenated with a single outcome or probability value to be transformed).

We first consider the case where $\mathbf{c}_1 = \mathbf{x}_A$, and $\mathbf{c}_2$ contain no information; that is, where the utility function is conditioned on all outcomes within a single gamble ("Intra-Gamble Outcome Context"):

$$V(A) = \sum_{i \in A} u(x_i, \mathbf{x}_A)\, \pi(p_i). \tag{8}$$

Notably, although $x_i \in \mathbf{x}_A$, which might seem redundant, the context vector $\mathbf{x}_A$ that is input to the utility function will not vary across evaluations of different x_i within the value computation for a gamble. Contained in this model class are disappointment aversion models (*55, 56*), which explicitly represent differences between outcomes as disappointment-related quantities of interest.

Next, we consider the case where $\mathbf{c}_1 = \{\mathbf{x}_A, \mathbf{x}_B\}$, and $\mathbf{c}_2$ contains no information; that is, where the utility function is conditioned on all outcomes across both gambles in a choice problem ("Inter-Gamble Outcome Context"):

$$V(A) = \sum_{i \in A} u(x_i, \mathbf{x}_A, \mathbf{x}_B)\, \pi(p_i). \tag{9}$$

Contained in this model class are regret aversion models (*57, 58*), which explicitly represent differences between outcomes across different gambles as regret-related quantities of interest.

We further consider the cases where additionally $\mathbf{c}_2 = \{\mathbf{p}_A, \mathbf{p}_B\}$ ("Inter-Gamble Outcome/Probability Context"):

$$V(A) = \sum_{i \in A} u(x_i, \mathbf{x}_A, \mathbf{x}_B)\, \pi(p_i, \mathbf{p}_A, \mathbf{p}_B), \tag{10}$$

although probabilities can also interact, and where $\mathbf{c}_1 = \mathbf{c}_2 = \{\mathbf{x}_A, \mathbf{x}_B, \mathbf{p}_A, \mathbf{p}_B\}$ ("Fully Contextual Multiplicative"):

$$V(A) = \sum_{i \in A} u(x_i, \mathbf{x}_A, \mathbf{x}_B, \mathbf{p}_A, \mathbf{p}_B)\, \pi(p_i, \mathbf{x}_A, \mathbf{x}_B, \mathbf{p}_A, \mathbf{p}_B), \tag{11}$$

where all possible information about both gambles conditions the subjective evaluation of all outcomes and probabilities.

Mixture of Theories models

We define a Mixture of Theories (MOT) model using separate convex combinations of utility and probability weighting functions:

$$V(A) = \sum_{i \in A} \left[\sum_j \omega_j \, u_j(x_i) \right] \left[\sum_k \omega_k \, \pi_k(p_i) \right] \tag{12}$$

We use a "mixture of experts" architecture (*37*): two neural networks take the contents of both gambles as input, and output weights $\{\omega_1, ..., \omega_j, ..., \omega_J\}$ and $\{\omega_1, ..., \omega_k, ..., \omega_K\}$ respectively using softmax output layers, such that $\sum_j \omega_j = 1$ and $\sum_k \omega_k = 1$. When weights are close to 0 or 1, the mixture implements a soft selection of one or the other subjective functions. When weights are close to 0.5, it implements functionally interpolated subjective functions. In practice, to reduce model parameters, the weights connecting the input layer and 20-unit hidden layer of both neural networks were shared, and only the output layers were separate. While any utility and probability weighting functions could be used, we use two "General Power" utility functions and two "Kahneman-Tversky" probability weighting function (see definitions in sections "Expected Utility models" and "Prospect Theory models"). Optimization of the parameters of all four functions, as well as the neural network that infers mixture weights on a per-problem basis, is done jointly, learning both the most useful subjective utility and probability weighting functions and their mixtures. In our experiments, one of the two probability weighting functions always approximated the identity function as a result of fitting the data, and thus for further sample efficiency we fix one of them to the identity function in advance when fitting final models. Finally, when one gamble dominates another (all outcomes are strictly better) a fixed probability (another jointly learned parameter, P_{dom}) is taken as the prediction of the overall model.

Best Estimate and Sampling Tools model

The Best Estimate and Sampling Tools (BEAST) model is different from many classic theories in that it does not employ subjective weighing functions, instead opting to explain decisions at a process level, specifically appealing to "mental sampling" (*27*). BEAST expresses the value of each gamble as the sum of the best estimate of its expected value and that of a set of sampling tools that correspond to four behavioral tendencies. Under this scheme, gamble A will be strictly preferred to gamble B if and only if:

$$[BEV_{\mathrm{A}} - BEV_{\mathrm{B}}] + [ST_{\mathrm{A}} - ST_{\mathrm{B}}] + e > 0, \tag{13}$$

where BEV_{A}–BEV_{B} is the advantage of gamble A over gamble B based on their expected values, ST_{A}–ST_{B} is the advantage based on alternative sampling tools, and e is a normal error term.

The sampling tools employed by BEAST include both biased and unbiased sample-based estimators designed to explain four psychological biases: (1) the tendency to assume the worst outcome of each gamble, (2) the tendency to weight all outcomes as equally likely, (3) sensitivity to the sign of the reward, and (4) the tendency to select the gamble that minimizes the probability of immediate regret. Further details can be found in (*27*).

When evaluating BEAST, we use the newest version proposed and implemented by the authors (*28*) and pre-fit to explain both classic choice problems and data from two choice competitions. We use the default parameters set by the authors since the model takes over a day to output predictions for our large dataset, since it was not designed specifically for scale, and thus takes considerable time to re-fit. Since BEAST was designed to produce predictions for five blocks of trials that have a very specific pattern of feedback, we used the best of these five predictions for each problem to prevent unfair comparison. When only a single block was used, performance was similar to Prospect Theory.

Relationships between models

When $u(\cdot)$ is the identity function $u(x) = x$, EU reduces to EV, and so EV is a form of EU. Further, when $\pi(p) = p$, PT reduces to EU. CPT cannot be reduced to PT, even when $\pi^{-}(\cdot) = \pi^{+}(\cdot)$, because the application of decision weights to outcomes depends on their sign, and because it cannot violate stochastic dominance. For these reasons, CPT is a unique subset of Value-Based theories, which also contains PT, EU, and EV as simple cases. Since a fully unconstrained neural network (Context-Dependent) can approximate nearly any function, it can approximate "Value-Based" functions, and is the most general class of models we consider. More constrained models that incorporate context are those in the Contextual Multiplicative class. When context vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ in these models contain no information, or are uninformative, this class reduces to Prospect Theory. The Contextual Multiplicative class contains MOT as a special case, *i.e.,* where conditioning of the psychophysical functions $u(\cdot)$ and $\pi(\cdot)$ in these models is restricted to the context-dependent selection of only two such unconditional forms (e.g., such as those in EU and PT).

Optimization

For all models above, we minimized the error function described with respect to all model parameters using gradient descent, and in particular the Adam (*59*) optimizer, which helps avoid having to conduct a large search of optimizer parameters themselves. To perform this optimization, we used the `jax` Python library (*60*) for automatic differentiation. In the case of structured models such as Neural PT, error is backpropagated through the decision function and further through the component functions or neural networks.

Out-of-sample model validation

Models with more parameters – especially neural networks, which can contain thousands or millions of parameters – can overfit data and often fail to generalize after fitting. To evaluate the generalization performance of models, we fit the parameters of each model using "training" data, and test them out-of-sample using unseen "test" data. In particular, we perform 50 splits of our data wherein a random 90% subset is used for training (or some subset of this overall training set), and the other 10% is used for testing. We also used this process to select the best hyperparameters of each model: the best of two initial learning rate values: 0.01, and 0.001 and the size of the hidden layers of each neural network (neural networks larger or smaller than those evaluated above did not yield better results in our tests).

Model analysis using Scientific Regret Minimization

Scientific Regret Minimization (*61*) enables researchers to leverage an unconstrained machine learning model to help build a more psychologically interpretable model by examining the residuals between the two models and iteratively incorporating more complex features. We first fit an expected utility function to the whole dataset (Fig. S6) and identified the largest residuals between this model and the unconstrained neural network. Many of these residuals were from "dominated gambles" – gambles in which one of the option's outcomes is greater than or equal to all the other option's outcomes. We then calculated the largest residuals after filtering out the dominated gambles. We estimated a Prospect Theory model for the gambles with residuals greater than 0.20, which resulted in a loss-averse utility function (Fig. S7) and an inverse-S-shaped probability weighting function (Fig. S8). This motivated the choice of components in the Mixture of Theories model presented in the main text.

Mixture of Theories Problem Analysis

The neural network component of our Mixture of Theories assigned weights greater than 0.5 for Utility Function 2 (UF 2) to 81% of the choice problems in our dataset, and for Probability Weighting Function 2 (PWF 2) to 12% of the choice problems in our dataset. Example choice problems with the highest mixture weights for UF 1, UF 2, PWF 1, and PWF 2 are shown in Tables 4, 5, 6, and 7, respectively. To understand the differences between the choice problems for which our mixture assigned different functions, we explored statistics of several choice problem descriptors that might explain this assignment. For example, we can compare the mean outcomes for choice problems with mixture weights above 0.5 to those with weights below 0.5. Visually apparent differences and individual-samples t-tests with unequal variances from Fig. S12 and S12 show meaningful differences for some descriptors but not others. Notably, context-dependent employment of different utility and probability weighting functions is not predicted by differences in expected value. Instead, choice of utility function was best predicted by minimum outcome, maximum outcome, and outcome variability across both gambles in a problem, while choice of probability weighting function was best predicted by minimum outcome and number of losses.

Mixture of Theories Ablation Analysis

Since MOT contains multiple components, we wanted to understand the relative importance of each in obtaining good predictive performance. To do this, we conducted an ablation analysis wherein we removed either the dominated-gambles component, one of the utility functions, or one of the probability weighting functions. For reference, we note that the original MOT model obtained an MSE score of 0.0113. The largest drop in performance (increase in error) came from removing one of the utility functions (MSE = 0.0132), suggesting that context effects relating to changes in utilities are most important. The next largest drop came from removing the dominated-gambles component (MSE = 0.0126). Finally, removing a probability weighting function, while least detrimental to performance, still had a sizable effect (MSE = 0.0121), indicating that context effects relating to changes in probability weighting are also important.

<u>Mixtures of Infinite Theories</u>

We also explored another class of models which we call Mixture of Infinite Theories (MOIT), where the parameters θ_h of a theory $h(\cdot)$ are inferred on the basis of each problem by a learned function $q(\cdot)$:

$$\theta_h = q(\mathbf{x}_A, \mathbf{x}_B, \mathbf{p}_A, \mathbf{p}_B), \tag{14}$$

such that the value of a gamble becomes

$$V(A) = h(\mathbf{x}_A, \mathbf{p}_A, \theta_h). \tag{15}$$

Notably, θ_h is allowed (and expected) to vary across and be inferred anew for each possible choice problem.

We consider in particular the case where the theory of interest is Prospect Theory:

$$V(A) = \sum_i u(x_i, \theta_1)\, \pi(p_i, \theta_2), \tag{16}$$

where $\theta_h = \{\theta_1, \theta_2\}$ contains the parameters for both the utility and weighting functions. In our implementation, $q(\cdot)$ is a neural network with its own parameters fit by gradient descent, and $u(\cdot)$ and $\pi(\cdot)$ are simple parametric forms that require only a handful of parameters (i.e., in θ_h) to be inferred. This model can be thought of as an alternative implementation of the Fully Contextually Multiplicative model, and also contains MOT as a special case, where only two possible sets of subjective function parameters can be assigned to each problem. Despite this, MOIT did not perform as well as MOT (MSE = 0.0155; see Table S1 for comparison), suggesting that it is too expressive to make efficient use of our data.

Replication Study

To allay concerns about data quality and platform-specific bias, we conducted a partial replication of our original dataset. We recruited 300 participants via Prolific (*62*), each of whom completed 300 trials, corresponding to five repetitions of 60 unique gambles randomly sampled from the set of 1,000 gambles with the most responses in the original dataset. Feedback was given for all decisions, unlike the original experiment in which feedback was given for only 80% of trials. All other aspects of this dataset match the original. The Pearson correlation between the choice probabilities for our original and replication data was $.80$, despite the difference in crowdsourcing platform, experiment structure, and problem distribution, as well as the intrinsic noise due to individual variability.

Models fit to the original dataset were then evaluated on this newly collected out-of-sample dataset. MSE scores for Neural EU, Neural PT, Value-Based, Mixture of Theories, and Context-Dependent models were 0.0197, 0.0165, 0.0123, 0.0110, and 0.0106 respectively, closely resembling the pattern of performance from our main analysis.

Prediction of Individual Behavior

Since our models are fit and evaluated on proportions aggregated across different individuals, it is possible that our models do not reflect the strategies that individuals actually employ. To rule out this possibility, we tested a subset of our models on out-of-sample individual responses from the replication study described in the previous section. Encouragingly, we found that Neural PT better predicted 63% of all individuals compared to Neural EU, confirmed by a one-tailed binomial test ($z = 4.35$, $p < 0.0001$), and our Context-Dependent neural network outperformed Neural PT for 74% of all individuals ($z = 8.08$, $p < 0.0001$). This indicates that better models of aggregate behavior in our case result in better prediction of a greater number of individuals. The absolute rates of any one model beating the other two were 17% for Neural

EU, 17% for Neural PT, and 66% for Context-Dependent. These results also suggest that, while individual differences clearly exist and may be substantial, there is also substantial overlap in the strategies individuals in our dataset employ.

Modeling individual differences directly with data-driven models can be challenging, because typically only small datasets can be collected per individual. To address this, we fine-tuned models that were previously already fitted to the aggregate data to data from individuals in our replication dataset. This allows us to start with a model of shared behavior, having learned from thousands of datapoints already, and then adapt that behavior uniquely for each individual given less data than would typically be needed. For each individual, we used 50 problems for fine-tuning, and the remaining 10 for testing. Models were fine-tuned for 500 gradient updates maximum. We fine-tuned MOT instead of Context-Dependent because it has interpretable components. To further improve sample efficiency, we used 7 units in the hidden layer of the mixture networks, which only marginally degrades performance in predicting aggregate data. We chose Prospect Theory as our baseline over which we want to improve performance, since it is a common choice for individual-level modeling. We found that fine-tuning MOT resulted in lower MSE (0.058) than fine-tuning Prospect Theory (MSE = 0.063), suggesting that models that better predict aggregate behavior may also be better starting points for modeling individuals. Fig S14 shows individual MOT fit, revealing variation in the utility and probability-weighting functions, as well as the relative tendency to employ some over others. Notably, when fine-tuning only the utility and probability weighting functions of MOT, but not the mixture networks, the advantage of MOT over PT was eliminated (MSE = 0.065), compared to when the mixture network was also fine-tuned as above, suggesting that context may be a more explanatory aspect of individual variation than subjective weighting. Lastly, when restricting inputs to the mixture network (which determine context) to outcomes only, we find identical performance to the original case (MSE = 0.058), suggesting that this aspect of context is largely invariant across individuals.

Interpreting Differences in Model Performance

Mean squared error (MSE) and crossentropy are both smooth error functions that allow for gradient-based optimization of model parameters, and performance metrics that are sensitive to the precise choice probabilities for each problem. However, the values themselves are not readily interpretable. Instead, the performance curves in our main analyses are perhaps most useful in illustrating the relative performance increases among models. For example, including context-dependence in models results in a performance increase many times larger than introducing global probability weighting (i.e., the leap from EU to PT).

Another way to measure performance is to assess the accuracy of each model: the proportion of problems where the model prediction for $P(A)$ is greater than 0.5 when the observed proportions are also greater than 0.5. Under this metric, the worst model we evaluated in our paper—the minimax heuristic (see Table 3)—obtained an accuracy score of 38.51%. Accuracy for the differentiable decision theories we evaluated ranged from 81.41% for Neural EU to 84.81% for Context-Dependent. However, this measure is completely insensitive to differences in the precise choice proportions, such that predictions of 0.51 and 0.99 are treated as equally correct, even though one may indicate almost indifferent behavior, and the other extreme confidence.

To give a better sense of model performance in predicting exact choice rates, we also calculated the proportion of model predictions within 0.05, 0.1, 0.2, and 0.3 of the observed choice rates. For Neural EU, these scores were 25.85%, 48.61%, 80.70%, and 95.27%. For Context-Dependent, they were 39.69%, 68.73%, 94.64%, and 99.64%.

Alternate Theories

We evaluated 21 additional theories from diverse modeling traditions spanning over five decades of various literatures, sampled from the ontology in (*63*). Results are shown in Table 3. This included nine heuristics, all of which performed worse than all other theories evaluated in our paper, even though the Priority Heuristic is in the complexity class motivated by our main analysis. Heuristics contain no model parameters except a single parameter to specify the probability of choosing the preferred gamble, as defined in (*63*), and optimized here using gradient descent. The rest of the theories we evaluated contained up to several parameters, as well as η, following the formulation for models in the main analysis specified above, all of which were optimized using gradient descent. This includes five "Counterfactual" models. Within this class are three forms of Disappointment Theory (*55, 56*), which fall into our own "Intra-Gamble Outcome Context" model class yet do not match its performance, and two forms of Regret Theory (*57, 58*), which fall into our own "Inter-Gamble Outcome Context" model class yet do not match its performance either. We also evaluated six "risk-as-value" models, which explicitly penalize risk and/or variability, and Transfer of Attention Exchange (TAX) (*45*), an influential recent model related to the family we evaluated in our paper ("Subjective Expected Utility"). Only the latter (TAX), out of all 21 theories in this section, was able to match the performance of Prospect Theory, although it falls into a more complex model class in our taxonomy ("Intra-Gamble Outcome Context"). Taken together, these results suggest that theories from the literature that fall into the classes we search may not be optimal exemplars of those classes, at least for our dataset. Further work will be needed to determine whether the psychological effects captured by these different models are more useful in a context-dependent manner.

Figs. S1 to S11

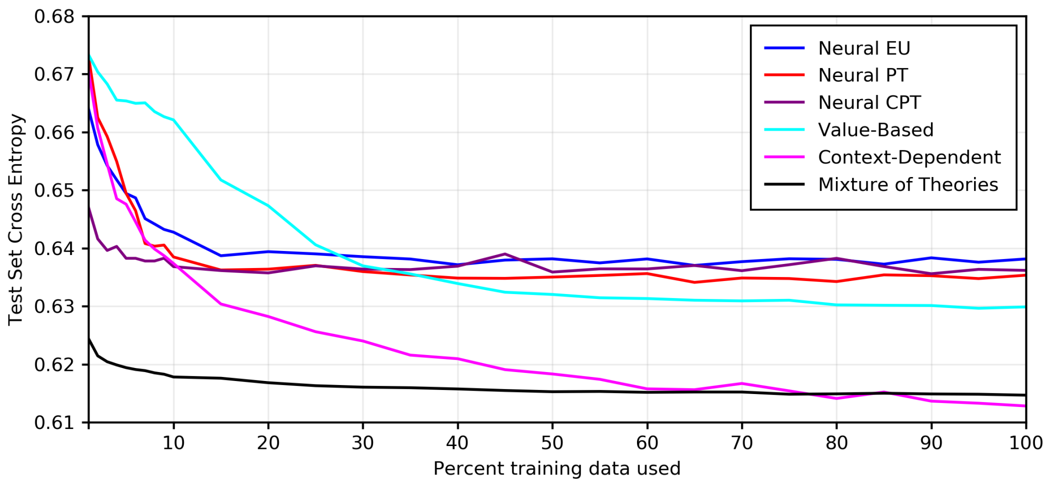

Fig. S1: Performance of differentiable decision theories using cross entropy error instead of mean squared error, which yields a similar overall pattern of results. Performance is assessed by model prediction error (cross entropy) on 1,000 unseen choice problems as a function of the amount of training data used for model fitting (approximately 9,000 choice problems, average of 50 runs).

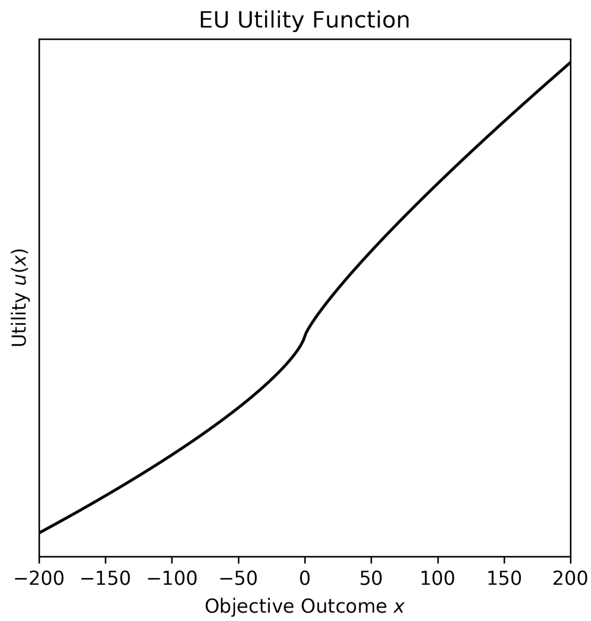

Fig. S2: EU utility function when fit to whole dataset.

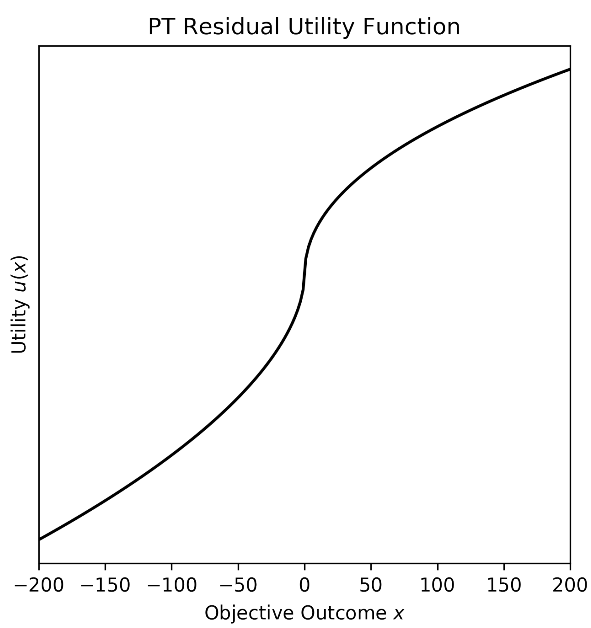

Fig. S3: PT utility function when fitting to largest residuals between EU and neural network.

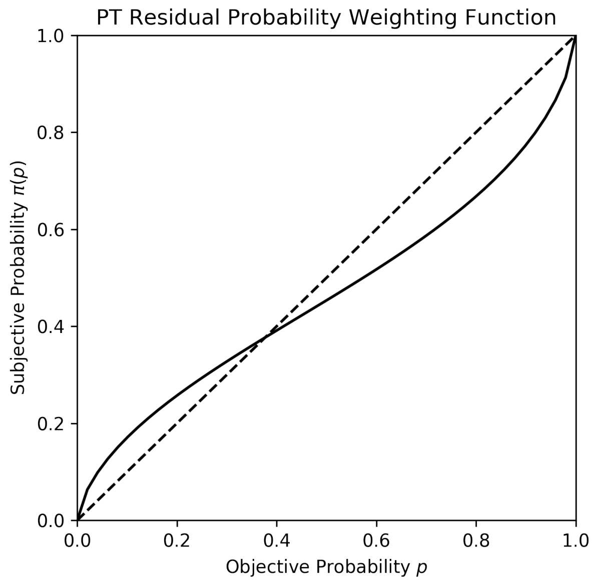

Fig. S4: PT probability weighting function when fitting to largest residuals between EU and neural network.

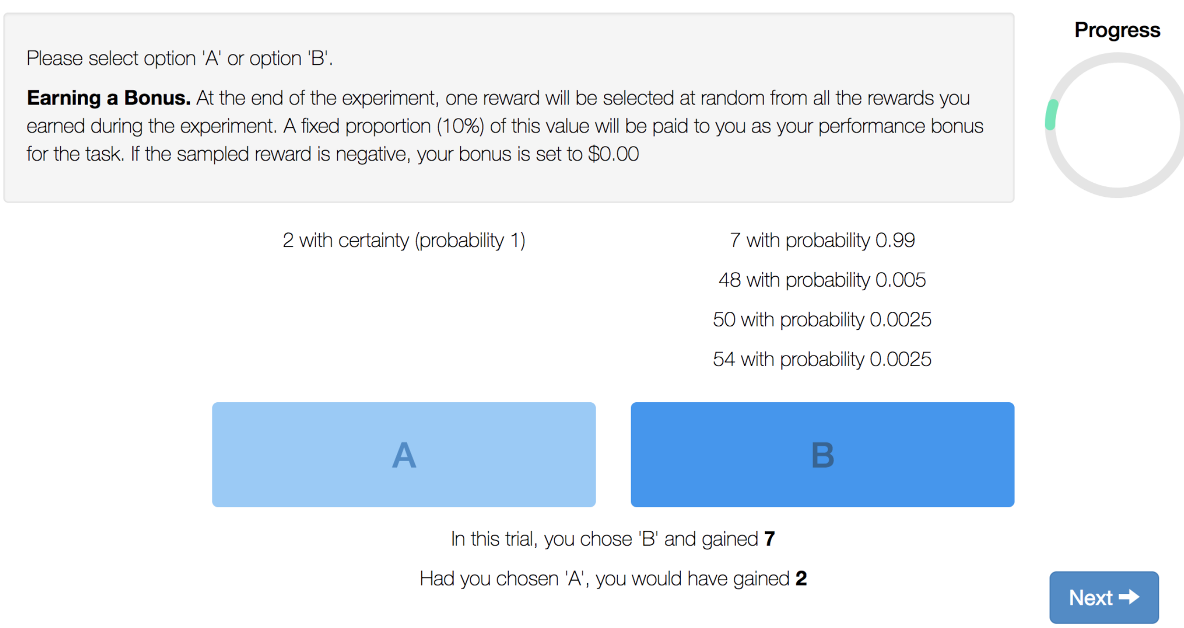

Fig. S5: Gamble selection interface for collecting human responses for the `choices13k` dataset.

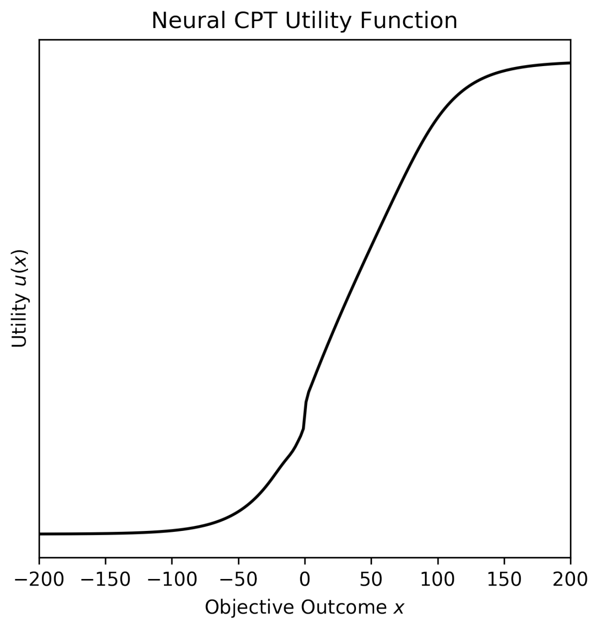

Fig. S6: Neural CPT utility function when fit to the whole dataset.

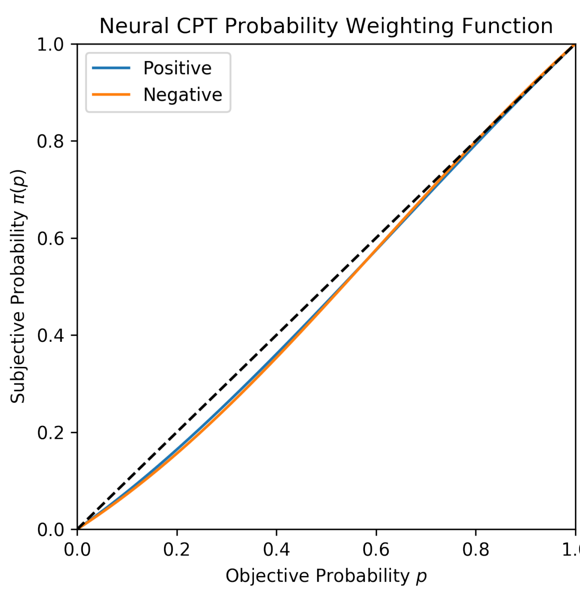

Fig. S7: Neural CPT probability weighting functions when fit to the whole dataset.

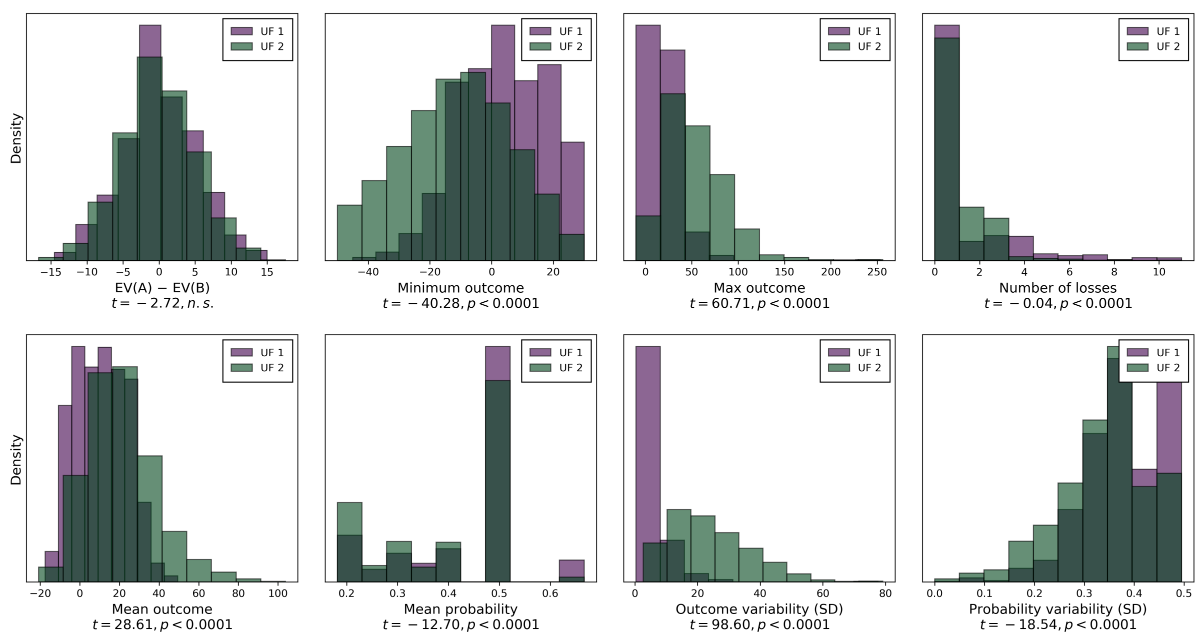

Fig. S8: Histograms of descriptors for similarities and differences among approximately 10,000 choice problems assigned by MOT (weights less than or greater than 0.5) to one of two utility functions. The descriptors that were most predictive of this assignment were the minimum outcome, the maximum outcome, and outcome variability across both gambles in a problem.

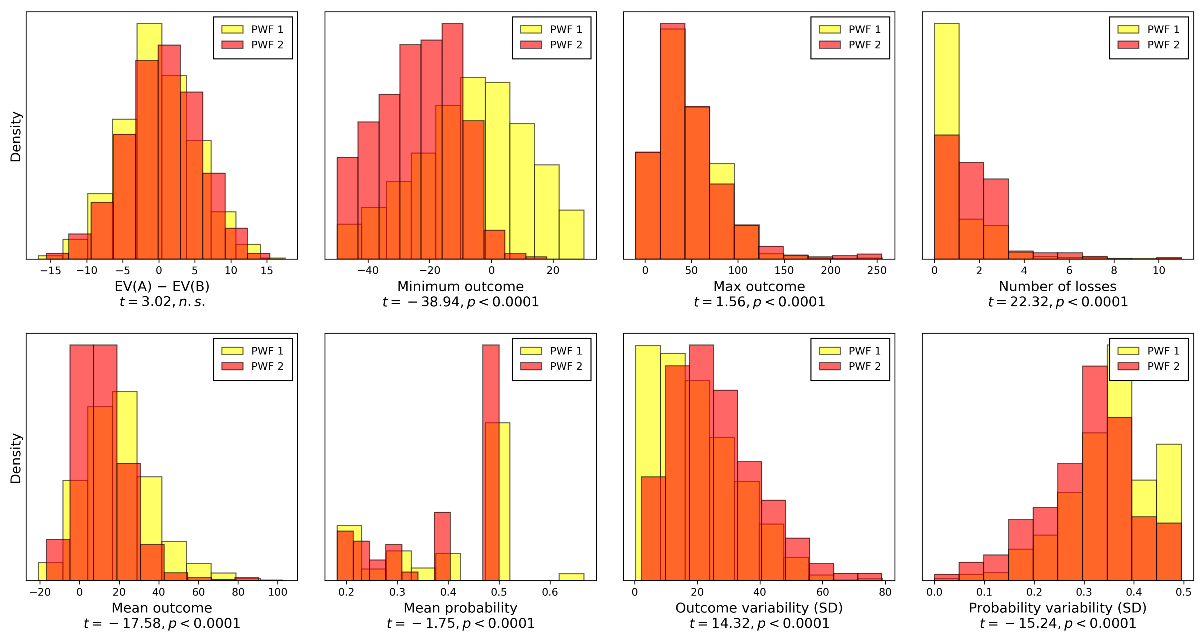

Fig. S9: Histograms of descriptors for similarities and differences among approximately 10,000 choice problems assigned by MOT (weights less than or greater than 0.5) to one of two probability weighting functions. The descriptors that were most predictive of this assignment were the minimum outcome and the number of losses across both gambles in a problem.

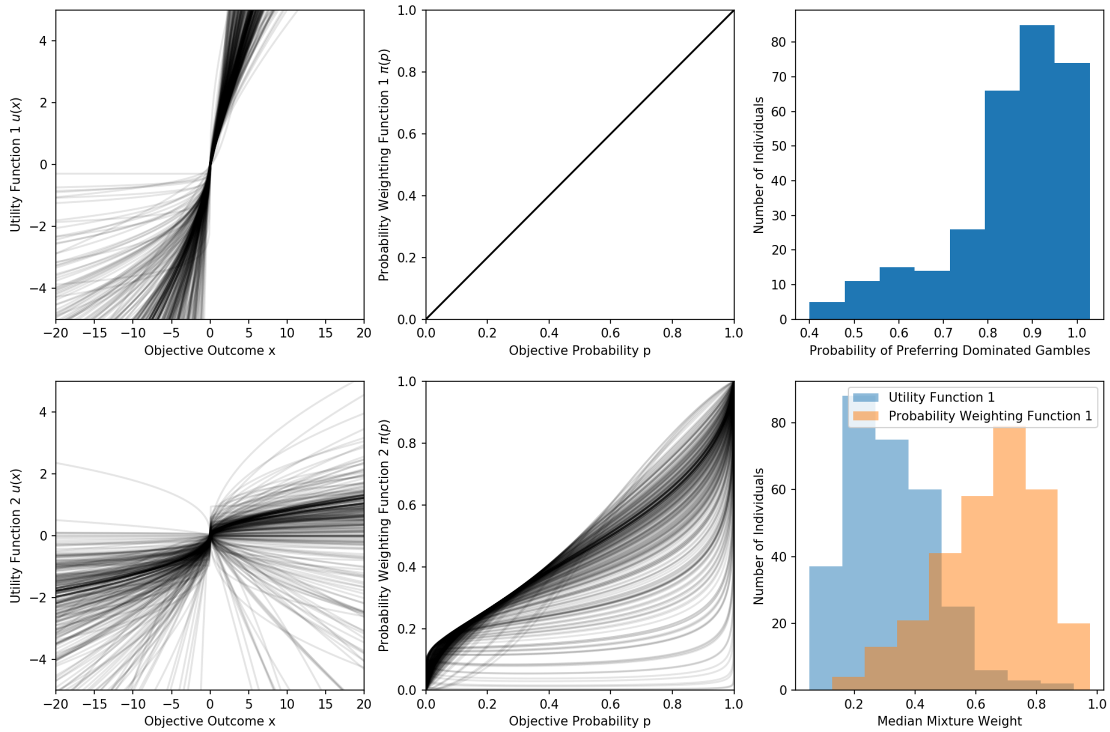

Fig. S10: Individual differences in MOT components fit to 300 different individuals.

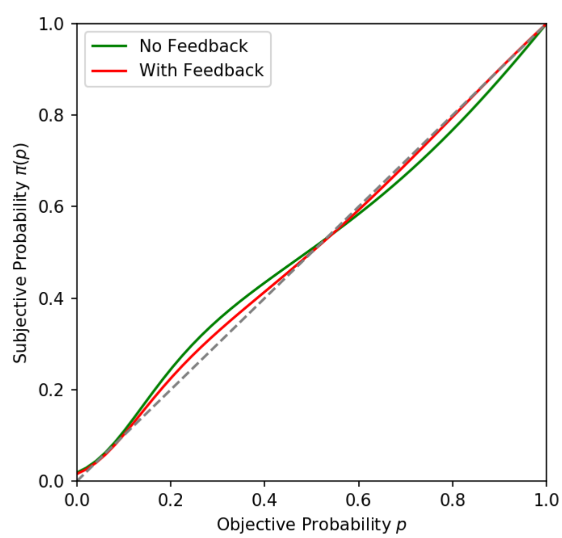

Fig. S11: Neural PT probability weighting functions when fitting to the subsets of our data where feedback was given to participants versus no feedback.

Tables S1 to S7

Model	Mean Square Error	Cross Entropy
Neural EU	.0217	.6382
Neural PT	.0204	.6354
Neural CPT	.0210	.6362
Value-Based	.0179	.6299
Context-Dependent	.0117	.6128
Mixture of Theories	.0113	.6147

Table S1: Performance of each differentiable decision theory as measured by both mean squared error and cross entropy when training on 90% of our data and testing on 10%. Scores are averaged across 50 random train/test data splits.

Table S2: Model results for different combinations of util-ity and probability weighting functions proposed by re-searchers. When the probability weighting function is the identity function, the model is EU. Otherwise, the model is PT.

Utility Function	Probability Weighting Function	MSE
Exp Loss Averse	Compound Invariance	0.0265
Exp Loss Averse	Constant Relative Sensitivity	0.0261
Exp Loss Averse	Exponential Power	0.0285
Exp Loss Averse	Hyperbolic Log	0.0274
Exp Loss Averse	Identity	0.0262
Exp Loss Averse	Kahneman-Tversky	0.0242
Exp Loss Averse	Log Odds Linear	0.0243
Exp Loss Averse	Neo Additive	0.0275
Exp Loss Averse	Power	0.0247
Exp Power Loss Averse	Compound Invariance	0.0351
Exp Power Loss Averse	Constant Relative Sensitivity	0.0378
Exp Power Loss Averse	Exponential Power	0.0346
Exp Power Loss Averse	Hyperbolic Log	0.0353
Exp Power Loss Averse	Identity	0.0374
Exp Power Loss Averse	Kahneman-Tversky	0.0361
Exp Power Loss Averse	Log Odds Linear	0.0353
Exp Power Loss Averse	Neo Additive	0.0346
Exp Power Loss Averse	Power	0.0358
General Linear Loss Averse	Compound Invariance	0.0252
General Linear Loss Averse	Constant Relative Sensitivity	0.0264
General Linear Loss Averse	Exponential Power	0.0257
General Linear Loss Averse	Hyperbolic Log	0.0256
General Linear Loss Averse	Identity	0.0268
General Linear Loss Averse	Kahneman-Tversky	0.0263
General Linear Loss Averse	Log Odds Linear	0.0255
General Linear Loss Averse	Neo Additive	0.0255
General Linear Loss Averse	Power	0.0264
General Power Loss Averse	Compound Invariance	0.0207
General Power Loss Averse	Constant Relative Sensitivity	0.0221
General Power Loss Averse	Exponential Power	0.0205
General Power Loss Averse	Hyperbolic Log	0.0207
General Power Loss Averse	Identity	0.0221
General Power Loss Averse	Kahneman-Tversky	0.0211
General Power Loss Averse	Log Odds Linear	0.0209

General Power Loss Averse	Neo Additive	0.0203
General Power Loss Averse	Power	0.0210
Identity	Compound Invariance	0.0256
Identity	Constant Relative Sensitivity	0.0264
Identity	Exponential Power	0.0256
Identity	Hyperbolic Log	0.0262
Identity	Identity	0.0266
Identity	Kahneman-Tversky	0.0268
Identity	Log Odds Linear	0.0259
Identity	Neo Additive	0.0261
Identity	Power	0.0266
Linear	Compound Invariance	0.0258
Linear	Constant Relative Sensitivity	0.0268
Linear	Exponential Power	0.0260
Linear	Hyperbolic Log	0.0258
Linear	Identity	0.0267
Linear	Kahneman-Tversky	0.0267
Linear	Log Odds Linear	0.0260
Linear	Neo Additive	0.0259
Linear	Power	0.0267
Linear Loss Averse	Compound Invariance	0.0253
Linear Loss Averse	Constant Relative Sensitivity	0.0267
Linear Loss Averse	Exponential Power	0.0253
Linear Loss Averse	Identity	0.0265
Linear Loss Averse	Kahneman-Tversky	0.0264
Linear Loss Averse	Log Odds Linear	0.0255
Linear Loss Averse	Neo Additive	0.0404
Linear Loss Averse	Power	0.0263
Log Loss Averse	Compound Invariance	0.0288
Log Loss Averse	Constant Relative Sensitivity	0.0317
Log Loss Averse	Exponential Power	0.0286
Log Loss Averse	Hyperbolic Log	0.0288
Log Loss Averse	Identity	0.0320
Log Loss Averse	Kahneman-Tversky	0.0299
Log Loss Averse	Log Odds Linear	0.0293
Log Loss Averse	Neo Additive	0.0282
Log Loss Averse	Power	0.0293
Norm Exp Loss Averse	Compound Invariance	0.0344
Norm Exp Loss Averse	Constant Relative Sensitivity	0.0374
Norm Exp Loss Averse	Exponential Power	0.0343
Norm Exp Loss Averse	Hyperbolic Log	0.0349

Norm Exp Loss Averse	Identity	0.0372
Norm Exp Loss Averse	Kahneman-Tversky	0.0362
Norm Exp Loss Averse	Log Odds Linear	0.0326
Norm Exp Loss Averse	Neo Additive	0.0347
Norm Exp Loss Averse	Power	0.0355
Norm Log Loss Averse	Compound Invariance	0.0299
Norm Log Loss Averse	Constant Relative Sensitivity	0.0300
Norm Log Loss Averse	Exponential Power	0.0294
Norm Log Loss Averse	Hyperbolic Log	0.0294
Norm Log Loss Averse	Identity	0.0304
Norm Log Loss Averse	Kahneman-Tversky	0.0301
Norm Log Loss Averse	Log Odds Linear	0.0296
Norm Log Loss Averse	Neo Additive	0.0290
Norm Log Loss Averse	Power	0.0292
Norm Power Loss Averse	Compound Invariance	0.0293
Norm Power Loss Averse	Constant Relative Sensitivity	0.0297
Norm Power Loss Averse	Exponential Power	0.0293
Norm Power Loss Averse	Hyperbolic Log	0.0296
Norm Power Loss Averse	Identity	0.0303
Norm Power Loss Averse	Kahneman-Tversky	0.0304
Norm Power Loss Averse	Log Odds Linear	0.0297
Norm Power Loss Averse	Neo Additive	0.0287
Norm Power Loss Averse	Power	0.0294
Power Loss Averse	Compound Invariance	0.0211
Power Loss Averse	Constant Relative Sensitivity	0.0223
Power Loss Averse	Exponential Power	0.0206
Power Loss Averse	Hyperbolic Log	0.0207
Power Loss Averse	Identity	0.0225
Power Loss Averse	Kahneman-Tversky	0.0209
Power Loss Averse	Log Odds Linear	0.0211
Power Loss Averse	Neo Additive	0.0205
Power Loss Averse	Power	0.0210
Quad Loss Averse	Compound Invariance	0.0870
Quad Loss Averse	Constant Relative Sensitivity	0.0384
Quad Loss Averse	Exponential Power	0.0930
Quad Loss Averse	Identity	0.0383
Quad Loss Averse	Log Odds Linear	0.0476
Quad Loss Averse	Neo Additive	0.0912
Quad Loss Averse	Power	0.0480

Decision Theory or Model	Category from (*63*)	MSE
Transfer of attention exchange (*45*)	Subjective Expected Utility	0.0207
Portfolio theory with variance (*64*)	Risk-as-value	0.0259
Portfolio theory with standard deviation (*65*)	Risk-as-value	0.0262
Below target model (*65*)	Risk-as-value	0.0251
Below-mean semivariance (*65*)	Risk-as-value	0.0252
Below-target semivariance (*65*)	Risk-as-value	0.0266
Coefficient of variation (*66*)	Risk-as-value	0.0266
Regret theory with expected value (*57*)	Counterfactual	0.0258
Regret theory with expected utility (*58*)	Counterfactual	0.0215
Disappointment theory with no rescaling (*55*)	Counterfactual	0.0266
Disappointment theory with expected value (*56*)	Counterfactual	0.0266
Disappointment theory with expected utility (*56*)	Counterfactual	0.0219
Better than average (*67*)	Heuristics	0.0447
Equiprobable (*67*)	Heuristics	0.0431
Low payoff elimination (*67*)	Heuristics	0.0423
Low expected payoff elimination (*67*)	Heuristics	0.0445
Least likely (*67*)	Heuristics	0.0488
Most likely (*67*)	Heuristics	0.0498
Minimax (*67*)	Heuristics	0.0499
Maximax (*67*)	Heuristics	0.0462
Priority heuristic (*68*)	Heuristics	0.0416

Table S3: Performance for a large selection of alternative decision theories fit to our entire dataset. Theories are organized according to the ontology in (*63*), and span over five decades: 1952 to 2008. No models perform better than our Neural PT model. Most models exhibit average fit, except for Transfer of attention exchange (*45*) which is competitive with Prospect Theory, and heuristics which largely fail altogether.

Gamble A	Gamble B	Weight
(0, 0.5), (2, 0.5)	(1, 1.0)	0.9825
(-1, 1.0), (-1, 0.0)	(-2, 0.5), (0, 0.5)	0.9875
(-2, 1.0), (-2, 0.0)	(-3, 0.5), (-1, 0.5)	0.9862
(6, 0.9), (7, 0.1)	(6, 1.0)	0.9847
(7, 0.01), (11, 0.99)	(11, 1.0)	0.9823
(2, 0.05), (4, 0.95)	(4, 1.0)	0.9874
(6, 0.2), (7, 0.8)	(7, 1.0)	0.9864
(1, 0.01), (2, 0.99)	(2, 1.0)	0.9884
(5, 0.1), (6, 0.9)	(6, 1.0)	0.987
(5, 0.99), (6, 0.01)	(5, 1.0)	0.9854

Table S4: Ten choice problems, formatted as lists of (outcome, probability) tuples, from our dataset that were assigned highest mixture weights to utility function 1 in our Mixture of Theo-ries model.

Gamble A	Gamble B	Weight
(-33, 0.6), (50, 0.4)	(-15, 0.5), (-2, 0.25), (0, 0.12), (4, 0.06), (12, 0.03), (28, 0.02), (60, 0.02)	0.9917
(10, 0.25), (27, 0.75)	(24, 0.95), (103, 0.02), (105, 0.01), (109, 0.01), (117, 0.0), (133, 0.0), (165, 0.0), (229, 0.0)	0.9905
(-6, 1.0), (-6, 0.0)	(-35, 0.75), (79, 0.12), (81, 0.06), (85, 0.03), (93, 0.02), (109, 0.01), (141, 0.0), (205, 0.0)	0.9912
(-11, 0.2), (-6, 0.8)	(-33, 0.8), (97, 0.1), (99, 0.05), (103, 0.02), (111, 0.01), (127, 0.01), (159, 0.0), (223, 0.0)	0.9917
(-16, 0.6), (92, 0.4)	(35, 0.9), (57, 0.01), (58, 0.04), (59, 0.04), (60, 0.01)	0.9922
(-20, 0.4), (57, 0.6)	(23, 0.99), (42, 0.0), (43, 0.0), (44, 0.0), (45, 0.0)	0.9905
(7, 0.99), (15, 0.01)	(2, 0.9), (91, 0.05), (93, 0.02), (97, 0.01), (105, 0.01), (121, 0.0), (153, 0.0), (217, 0.0)	0.991
(-30, 0.5), (54, 0.5)	(-10, 0.25), (9, 0.38), (11, 0.19), (15, 0.09), (23, 0.05), (39, 0.05)	0.9907
(-6, 1.0), (-6, 0.0)	(-13, 0.9), (110, 0.05), (112, 0.02), (116, 0.01), (124, 0.01), (140, 0.0), (172, 0.0), (236, 0.0)	0.9916
(0, 1.0), (0, 0.0)	(-9, 0.8), (65, 0.1), (67, 0.05), (71, 0.02), (79, 0.01), (95, 0.01), (127, 0.0), (191, 0.0)	0.9908

Table S5: Ten choice problems, formatted as lists of (outcome, probability) tuples, from our dataset that were assigned highest mixture weights to utility function 2 in our Mixture of Theo-ries model.

Gamble A	Gamble B	Weight
(14, 0.6), (44, 0.4)	(26, 1.0)	0.999
(4, 0.75), (81, 0.25)	(23, 1.0)	0.9991
(5, 0.75), (62, 0.25)	(19, 1.0)	0.9992
(7, 0.6), (52, 0.4)	(25, 1.0)	0.9995
(2, 0.5), (44, 0.5)	(23, 1.0)	0.999
(15, 0.9), (88, 0.1)	(22, 1.0)	0.9991
(6, 0.8), (66, 0.2)	(18, 1.0)	0.9991
(17, 0.75), (61, 0.25)	(28, 1.0)	0.9992
(14, 0.9), (108, 0.1)	(23, 1.0)	0.9991
(5, 1.0), (5, 0.0)	(-4, 0.95), (-1, 0.02), (1, 0.01), (5, 0.01)	0.999

Table S6: Ten choice problems, formatted as lists of (outcome, probability) tuples, from our dataset that were assigned highest mixture weights to probability weighting function 1 in our Mixture of Theories model.

Gamble A	Gamble B	Weight
(-6, 0.75), (54, 0.25)	(-46, 0.25), (28, 0.75)	0.9479
(-20, 0.6), (8, 0.4)	(-43, 0.2), (-10, 0.4), (-8, 0.2), (-4, 0.1), (4, 0.05), (20, 0.02), (52, 0.02)	0.9552
(-42, 0.5), (102, 0.5)	(-49, 0.25), (53, 0.75)	0.9496
(-23, 0.5), (73, 0.5)	(-39, 0.25), (49, 0.75)	0.9486
(-13, 0.95), (107, 0.05)	(-44, 0.6), (49, 0.4)	0.9488
(-6, 0.9), (54, 0.1)	(-42, 0.4), (43, 0.6)	0.9483
(-14, 0.8), (64, 0.2)	(-45, 0.4), (49, 0.6)	0.9493
(-12, 0.9), (67, 0.1)	(-50, 0.5), (37, 0.5)	0.9494
(-23, 0.6), (54, 0.4)	(-42, 0.4), (35, 0.6)	0.9484
(-10, 1.0), (-10, 0.0)	(-33, 0.1), (-19, 0.22), (-15, 0.22), (-13, 0.45)	0.9627

Table S7: Ten choice problems, formatted as lists of (outcome, probability) tuples, from our dataset that were assigned highest mixture weights to probability weighting function 2 in our Mixture of Theories model.

References and Notes

1. N. C. Barberis, Thirty years of prospect theory in economics: A review and assessment. *J. Econ. Perspect.* **27**, 173–196 (2013). doi:10.1257/jep.27.1.173

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, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. S. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, M. Wellman, Machine behaviour. *Nature* **568**, 477–486 (2019). doi:10.1038/s41586-019-1138-y Medline

6. D. Bernoulli, Exposition of a new theory on the management of risk. *Econometrica* **22**, 23–36 (1954). doi:10.2307/1909829

7. L. J. Savage, *The Foundations of Statistics* (Courier, 1972).

8. D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–292 (1979). doi:10.2307/1914185

9. M. Allais, Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine. *Econometrica* **21**, 503–546 (1953). doi:10.2307/1907921

10. H. J. Einhorn, R. M. Hogarth, Decision making under ambiguity. *J. Bus.* **59** (S4), S225–S250 (1986). doi:10.1086/296364

11. D. Ellsberg, Risk, ambiguity, and the savage axioms. *Q. J. Econ.* **75**, 643–669 (1961). doi:10.2307/1884324

12. A. Tversky, D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992). doi:10.1007/BF00122574

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, Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017). doi:10.1177/1745691617693393 Medline

16. A. Peysakhovich, J. Naecker, Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *J. Econ. Behav. Organ.* **133**, 373–384 (2017). doi:10.1016/j.jebo.2016.08.017

17. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018). Medline

18. J. S. Hartford, J. R. Wright, K. Leyton-Brown, *Adv. Neural Inf. Process. Syst.* **29**, 2424–2432 (2016). 10.14288/1.0319323

19. L. He, P. P. Pantelis, S. Bhatia, The wisdom of model crowds. *Manage. Sci.*, in press.

20. J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016). doi:10.1126/science.aaf2654 Medline

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, The theory of decision making. *Psychol. Bull.* **51**, 380–417 (1954). doi:10.1037/h0053870 Medline

24. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015). doi:10.1038/nature14539 Medline

25. G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Contr. Signals Syst.* **2**, 303–314 (1989). doi:10.1007/BF02551274

26. K. Hornik, Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991). doi:10.1016/0893-6080(91)90009-T

27. I. Erev, E. Ert, O. Plonsky, D. Cohen, O. Cohen, From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychol. Rev.* **124**, 369–409 (2017). doi:10.1037/rev0000062 Medline

28. O. Plonsky, R. Apel, E. Ert, M. Tennenholtz, D. Bourgin, J. C. Peterson, D. Reichman, T. L. Griffiths, S. J. Russell, E. C. Carter, J. F. Cavanagh, I. Erev, 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, Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85–117 (2015). doi:10.1016/j.neunet.2014.09.003 Medline

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, Original and cumulative prospect theory: A discussion of empirical differences. *J. Behav. Decis. Making* **10**, 53–64 (1997). doi:10.1002/(SICI)1099-0771(199703)10:1<53:AID-BDM245>3.0.CO;2-1

36. K. Mukherjee, A dual system model of preferences under risk. *Psychol. Rev.* **117**, 243–255 (2010). doi:10.1037/a0017884 Medline

37. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Adaptive mixtures of local experts. *Neural Comput.* **3**, 79–87 (1991). doi:10.1162/neco.1991.3.1.79 Medline

38. N. Stewart, N. Chater, G. D. Brown, Decision by sampling. *Cognit. Psychol.* **53**, 1–26 (2006). doi:10.1016/j.cogpsych.2005.10.003 Medline

39. R. Bhui, S. J. Gershman, Decision by sampling implements efficient coding of psychoeconomic functions. *Psychol. Rev.* **125**, 985–1001 (2018). doi:10.1037/rev0000123 Medline

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. doi:10.1109/CVPR.2009.5206848

41. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012). doi:10.1145/3065386

42. T. L. Griffiths, Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015). doi:10.1016/j.cognition.2014.11.026 Medline

43. I. Gilboa, Expected utility with purely subjective non-additive probabilities. *J. Math. Econ.* **16**, 65–88 (1987). doi:10.1016/0304-4068(87)90022-X

44. D. Schmeidler, Subjective probability and expected utility without additivity. *Econometrica* **57**, 571 (1989). doi:10.2307/1911053

45. M. H. Birnbaum, New paradoxes of risky decision making. *Psychol. Rev.* **115**, 463–501 (2008). doi:10.1037/0033-295X.115.2.463 Medline

46. M. Scholten, D. Read, Prospect theory and the “forgotten” fourfold pattern of risk preferences. *J. Risk Uncertain.* **48**, 67–83 (2014). doi:10.1007/s11166-014-9183-2

47. A. Saha, Expo-power utility: A ‘flexible’ form for absolute and relative risk aversion. *Am. J. Agric. Econ.* **75**, 905–913 (1993). doi:10.2307/1243978

48. D. Peel, J. Zhang, The expo-power value function as a candidate for the work-horse specification in parametric versions of cumulative prospect theory. *Econ. Lett.* **105**, 326–329 (2009). doi:10.1016/j.econlet.2009.09.007

49. D. Prelec, The probability weighting function. *Econometrica* **66**, 497 (1998). doi:10.2307/2998573

50. R. Gonzalez, G. Wu, On the shape of the probability weighting function. *Cognit. Psychol.* **38**, 129–166 (1999). doi:10.1006/cogp.1998.0710 Medline

51. H. P. Stott, Cumulative prospect theory's functional menagerie. *J. Risk Uncertain.* **32**, 101–130 (2006). doi:10.1007/s11166-006-8289-6

52. R. D. Luce, Reduction invariance and Prelec's weighting functions. *J. Math. Psychol.* **45**, 167–179 (2001). doi:10.1006/jmps.1999.1301 Medline

53. A. al-Nowaihi, S. Dhami, A simple derivation of Prelec's probability weighting function. *J. Math. Psychol.* **50**, 521–524 (2006). doi:10.1016/j.jmp.2006.07.006

54. M. Abdellaoui, O. l'Haridon, H. Zank, Separating curvature and elevation: A parametric probability weighting function. *J. Risk Uncertain.* **41**, 39–65 (2010). doi:10.1007/s11166-010-9097-6

55. D. E. Bell, Disappointment in decision making under uncertainty. *Oper. Res.* **33**, 1–27 (1985). doi:10.1287/opre.33.1.1

56. G. Loomes, R. Sugden, Disappointment and dynamic consistency in choice under uncertainty. *Rev. Econ. Stud.* **53**, 271 (1986). doi:10.2307/2297651

57. D. E. Bell, Regret in decision making under uncertainty. *Oper. Res.* **30**, 961–981 (1982). doi:10.1287/opre.30.5.961

58. G. Loomes, R. Sugden, Regret theory: An alternative theory of rational choice under uncertainty. *Econ. J. (Lond.)* **92**, 805 (1982). doi:10.2307/2232669

59. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (22 December 2014).

60. J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, S. Wanderman-Milne, "JAX: composable transformations of Python+NumPy programs" (GitHub, 2018); https://github.com/google/jax.

61. M. Agrawal, J. C. Peterson, T. L. Griffiths, Scaling up psychology via scientific regret minimization. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8825–8835 (2020). doi:10.1073/pnas.1915841117 Medline

62. S. Palan, C. Schitter, Prolific.ac—A subject pool for online experiments. *J. Behav. Exp. Finance* **17**, 22–27 (2018). doi:10.1016/j.jbef.2017.12.004

63. L. He, W. J. Zhao, S. Bhatia, An ontology of decision models. *Psychol. Rev.* (2020). doi:10.1037/rev0000231

64. H. Markowitz, Portfolio selection. *J. Finance* **7**, 77–91 (1952). doi:10.1111/j.1540-6261.1952.tb01525.x

65. P. C. Fishburn, Mean-risk analysis with risk associated with below-target returns. *Am. Econ. Rev.* **67**, 116 (1977).

66. E. U. Weber, S. Shafir, A.-R. Blais, Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychol. Rev.* **111**, 430–445 (2004). doi:10.1037/0033-295X.111.2.430 Medline

67. W. Thorngate, Efficient decision heuristics. *Behav. Sci.* **25**, 219–225 (1980). doi:10.1002/bs.3830250306

68. E. Brandstätter, G. Gigerenzer, R. Hertwig, The priority heuristic: Making choices without trade-offs. *Psychol. Rev.* **113**, 409–432 (2006). doi:10.1037/0033-295X.113.2.409 Medline