Stochastic satisficing account of choice and confidence in uncertain value-based decisions

Uri Hertz[1,2]*, Bahador Bahrami[2], Mehdi Keramati[3]

[1]School of Advanced Study, University of London,

[2]Institute of Cognitive Neuroscience, University College London

[3]The Gatsby Computational Neuroscience Unit, University College London

*Correspondence: Uri Hertz, u.hertz@ucl.ac.uk, UCL Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, UK, +44-20-76795429

Short title:

Choice and confidence in uncertain decisions

Keywords:

Decision making, bounded rationality, uncertainty, reinforcement learning, confidence, metacognition

Introductory paragraph

Every day we make choices under uncertainty; choosing what route to work or which queue in a supermarket to take, for example. It is unclear how outcome variance, e.g. waiting time in a queue, affects decisions when outcome is stochastic and continuous. For example, how does one choose between an option with unreliable but high expected reward, and an option with more certain but lower expected reward? Here we used an experimental design where two choices' payoffs took continuous values, to examine the effect of outcome variance on decisions and confidence. Inconsistent with expected utility predictions, our participants' probability of choosing the good option decreased when both better and worse options' payoffs were more variable. Confidence ratings were affected by outcome variability only when choosing the good option. Inspired by the satisficing heuristic, we propose a "stochastic satisficing" (SSAT) model for choosing between options with continuous uncertain outcomes. In this model, decisions are made by comparing the available options' probability of exceeding an acceptability threshold and confidence reports scale with the chosen option's satisficing probability. The SSAT model best explained choice behaviour and most successfully predicted confidence ratings. We further tested the model's prediction in a second experiment where choice and confidence behaviours were found to be consistent with the SSAT simulations. Our model and experimental results generalize the cognitive heuristic of satisficing to stochastic contexts and thus provide an account of bounded rationality in the face of uncertainty.

Introduction

Every morning most people have to pick a route to work. While the shortest route may be consistently busy, others may be more variable, changing from day to day. The choice of which route to take impacts the commuting time and is ridden with uncertainty. Decision making under uncertainty has been studied extensively using scenarios with uncertain rewards [5–7]. In such scenarios, participants choose between multiple lotteries where each lottery can lead to one of the two (or several) consequences with different probabilities. Standard models like expected utility theory [3] and prospect theory [4] suggest parsimonious formulations for how the statistics of such binomial (or multinomial) distributions of outcomes determine the value (otherwise known as utility) of a lottery. These models, for example, explain the fact that in certain ranges people prefer certain small rewards to bigger more uncertain ones [1,2].

However, the commuting problem described here highlights the pervasive but much less studied relevance of outcome variance to decisions with continuous (rather than binary) outcomes. It is not straightforward how one's choice of the route could be decided using the heuristics applicable to binary win/lose outcomes.

Early studies of bounded rationality [8,9] introduced the concept of satisficing according to which, individuals replace the computationally expensive question "which is my best choice?" with the simpler and most-of-the-times adequately beneficial question "which option is good enough?". More precisely, instead of finding the best solution, decision makers settle for an option that satisfices an acceptability threshold [8]. In the case of commuting, such acceptability threshold could be "the latest time one affords to arrive at work". Here we generalize the satisficing theory to decision-making under uncertainty. Our "stochastic satisficing" model suggests that when outcomes are uncertain, one could evaluate -with reasonably simple and general assumptions about the probability distributions of outcomes- the available options' *probability* of exceeding an acceptability threshold, and choose between the options by comparing their satisficing probability [10]. Our commuter's stochastic satisficing heuristic could then be expressed as "which route is *more likely* to get me to work before X o'clock?"

The effect of uncertainty on confidence reports is commonly studied in perceptual detection tasks where one has to detect a world state from noisy stimuli (e.g. dots are moving to the left or to the right) [11–14]. Sanders and colleagues (2016) argued that confidence reports in perceptual decisions relates to the Bayesian formulation of confidence used in hypothesis testing. In this view, subjective confidence conveys the posterior probability that an uncertain choice is correct, given the agent's prior knowledge, and noisy input information. Following this scheme and generalizing it to value-based contexts, our probabilistic satisficing heuristic is naturally fit to account for the computational underpinnings of choice confidence and draws strong predictions about how confidence would vary with outcome variance. In fact, if choices were made by the probabilistic satisficing heuristic described above, confidence in those choices would be directly proportional to the probability that the chosen option exceeded the acceptability threshold. A choice whose probability of exceeding the acceptability threshold, is higher should be made more confidently than another that barely passes the criterion, even if they have equal expected values.

Here we asked if, and how, human decision makers learn and factor outcome variance in their choices between options with independent continuous returns. We hypothesised that decision makers use a stochastic satisficing heuristic to make their choices and that their confidence conveys the estimated probability of the chosen option's value to exceed the satisficing criterion. In two experiments, we used a two-armed bandit task in which the expected values and variances associated with outcomes of each arm were systematically manipulated. We tested the stochastic satisficing model against expected reward model, risk sensitive model [1] and an expected utility model [15], that propose alternative ways of computing choice and confidence as a function of the estimated statistics of options' returns.

In experiment 1, we compared the models based on their fit to the choice behaviour. We then empirically examined the models' qualitatively different predictions for confidence against the behavioural data. In experiment 2, we designed a new payoff structure for the two-arm bandit and ran simulations of the competing models using the best-fit parameter values extracted from experiment 1. These simulations gave us qualitative predictions for choice and confidence by each model for Experiment 2. We compared these predictions to the empirical data obtained from a new set of participants. The results of both experiments strongly favoured the use of the stochastic satisficing heuristic in choices between options with continuous outcomes.

Results

Participants performed a two-armed bandit task online where rewards were hidden behind two doors (Fig. 1A) and the reward magnitudes followed different probability distributions (Fig. 1B). On each trial, the participant decided which door to open, and expressed their choice confidence using a combined choice-confidence scale. Choosing the left side of the scale indicated choice of the left door and distance from the midline (ranged between 1: uncertain, to 6: certain) indicated the choice confidence. After the decision, the reward behind the chosen door was revealed and a new trial started. Each experimental condition was devised for a whole block of consecutive trials during which the parameters (mean and variance) governing the reward distribution for each door were held constant. Each block lasted between 27 to 35 trials. Transition from one block to the next happened seamlessly and participants were not informed about the onset of a new block.

In our first experiment, within each condition, the reward behind one door was drawn from a distribution with a higher mean (65, i.e. the "good" option) than the other door (35, i.e. the "bad" option). The variances of the bad and good options could independently be high ($H=25^2=625$) or low ($L=10^2=100$), resulting in a 2x2 design comprising 4 experimental conditions: 'L-L','L-H','H-L' and 'H-H'. In this notation, the first and the second letters indicate the variances of the bad and good options, respectively (Fig. 1B).

Experiment 1

Participants' trial-by-trial probability of choosing the good option, in each condition, started at chance level and increased until it reached a stable level after about 10 trials (N=65, Fig. 1C, left). To assess the asymptotic level of performance, within each participant, we averaged the probability of choosing the good option between trials 10 to 25 in each experimental condition. Probability of choosing the good option was highest in the 'L-L' and lowest during the 'H-H' condition (Fig. 1C, right). A repeated measure ANOVA test with the variances of the good and bad options as within-subject factors was used to evaluate this pattern. The effects of both variance factors were significant (variance of good option: $F(1,194)=22.24$, $p=0.00001$, variance of bad option: $F(1,194)=5.2$, $p=0.026$). This result

indicates an asymmetric effect of outcomes' payoff variances on choice: increased variance of the good option reduced the probability of choosing the good option, whereas increased variance of the bad option increased the probability of choosing the bad option. This variance-dependent choice pattern demonstrates that decision-making depended not only on the expected rewards, but also on their variances: increasing the variance decreased choice accuracy.

Stochastic satisficing account of choice

To examine the use of a probabilistic satisficing heuristic and acceptability threshold in decisions under uncertainty we devised the stochastic satisficing (SSAT) heuristic model (See Methods for the formal description) in which the mean and variance of rewards obtained by each of the two doors are tracked in a trial-by-trial manner. In this model, decision is made by comparing the probability of each option yielding a reward above an acceptability threshold, i.e. being good enough. In fact, given the estimated probability distribution over the rewarding outcome of a choice, the model computes the total mass under this distribution that is above the acceptability threshold (Fig. 2A). This cumulative quantity is then used to determine the probability of choosing that option. Such mechanism can capture the asymmetric effect of payoff variance on choice, as better option (i.e. higher than threshold) becomes less likely to exceed the acceptability threshold as its variance increases, while bad option (below threshold) becomes more likely to exceed the threshold as their variance increase. Upon making the choice and receiving reward feedback from the environment, the model updates the distribution over the value of the chosen action. The value of the unchosen action(s), however, drifts toward the acceptability threshold, modelling a gradual forgetting effect of the unexplored events (i.e. the probability of crossing the threshold converges to 50% for the unchosen option).

We compared this model's fit to the choice data with three alternative decision models that were previously used in similar settings (see SI Appendix for the details of the competing models). The first was a classical expected reward ('Reward') model that tracks the expected reward from each door on every time-step [16,17]. Choice is then made according to the expected reward of each option [3]. The second was the 'Risk Sensitive Temporal Differencing" ('Risk') model suggested by Niv et al. [1], which was successful at capturing risk aversion when rewards were uncertain. This model employed separate learning rates for positive versus negative prediction errors. Third, we examined an expected utility model ('Utility'), fitting a utility function to the rewards, and effectively penalizing options for payoff's variance [3,15].

We fitted the four models to all the choices made by participants (240 trials per participant) using Monte-Carlo-Markov-Chain (MCMC) procedure [18]. After correction for the number of parameters using Watanabe-Akaike information criterion (WAIC) [19], we compared posterior likelihood estimates obtained for each participant, for each model (Fig. S1). The SSAT model

gave the best fit to the choices, its WAIC scoring on average 10 points less than that of the competing models (pairwise t-tests: SSAT vs. 'Reward': t(64)=3.55, p=0.0007; SSAT vs. 'Risk': t(64)=3.2, p=0.002; SSAT vs. 'Utility': t(64)=3.89, p=0.0002).

Furthermore, to test how our model's output corresponded to the patterns of choices during the stable phase of experimental conditions (Fig. 1C, right), we extracted the trial-by-trial probability of choosing the good option estimated by the four models and averaged those probabilities between the 10^{th} and 25^{th} trials of each experimental condition (Fig. S2A). Critically, the SSAT model was the only one that captured the qualitative pattern of the effect of payoff variance on choice, with gradual increase in preferring the good option in the same order (L-L > H-L > L-H > H-H) found in the behavioural data (Fig. 2B). To quantify this observation we regressed each participant's averaged model-estimated probabilities, from his/her average choices in each experimental condition (see SI Appendix). We therefore obtained the goodness of fit (R^2) for each participant. The SSAT model gave the best fit to the pattern of choices compared to other models (mean±SEM: SSAT R^2=0.64±0.04, 'Reward' R^2=0.48±0.04, 'Risk' R^2=0.5±0.04, 'Utility' R^2=0.48±0.04, paired t-test: SSAT vs. 'Reward': t(64)=3.2; p=0.002; SSAT vs. 'Risk': t(64)=2.7; p=0.008) and SSAT vs. 'Utility' model t(64)=3 p=0.003).

Model predictions and behavioural data for confidence ratings

We hypothesize that confidence in choice reflects the subjective probability that the value of the chosen option exceeded the acceptability threshold (i.e., the total mass under the value distribution of the chosen option that is more than the acceptability threshold). To test this hypothesis, we simulated the models, with their individual free parameters fitted to trial-by-trial choice data, to draw predictions for the confidence reports. Following previous studies that examined confidence in value-based decisions [20,21], for the 'Reward', 'Risk' and 'Utility' models, confidence was defined as proportional to the estimated decision variable: means of rewards for 'Reward' and 'Risk' model and expected utility options for the 'Utility' model. Focusing on trials 10-25 of each experimental condition, we calculated the average confidence for each model's simulation when choosing the good option and when choosing the bad option in each condition. All models predicted lower confidence when choosing the bad, as compared to the good, option (Fig. S2B). Additionally, all models predicted similar confidence levels across conditions when the bad option was chosen. When choosing the good option, however, the SSAT model's prediction (Fig. 2B, middle panel) was the only one consistent with the behaviour (Fig. 2C, middle panel). SSAT model predicted higher confidence when variance of the good option was low (i.e. 'L-L','H-L'). The other models did not predict a variance effect on confidence (Fig. S2B).

Using a repeated measures ANOVA we found that the main effect of the good, but not the bad, option's variance was significant when choosing the good option (good option variance: F(1,194)=33.32, p<0.00001 , bad option variance: F(1, 194)=0.02, p=0.89). When

choosing the bad option, confidence ratings were generally lower (paired t-test t(64)=8.3, p=9e-12) and were not significantly different across experimental conditions. This main effect of good option's variance on confidence was only predicted by the SSAT model.

To quantitatively evaluate the models' prediction of the confidence-report pattern, we regressed the averaged estimated confidence ratings obtained from the four models, from the participants' confidence ratings and obtained the goodness of fit (R^2) for each participant (see Methods). The SSAT model gave the best fit to confidence reports (mean±SEM: SSAT R^2=0.4±0.04, 'Reward' R^2=0.34±0.036, 'Risk' R^2=0.35±0.035, 'Utility' R^2=0.35±0.037, paired t-test: SSAT vs. 'Reward': t(64)=4.25; p=0.00008; SSAT vs. 'Risk': t(64)=2.68; p=0.01) and SSAT vs. 'Utility' model (t(64)=2p=0.04) (See Fig. S2B).

Compared to the competing models, the stochastic satisficing model had better success at explaining choices and confidence patterns in Experiment 1, both qualitatively and quantitatively. A counterintuitive prediction of the model borne out by the behavioural data was the difference between the two conditions involving unequal variances (i.e. L-H and H-L conditions). Stochastic satisficing predicted –and the data confirmed– a difference in confidence (compare L-H and H-L in Fig. 2B, C) despite identical expected values for the chosen (good) option in these two conditions. In Experiment 2, we focused on the choice between options with unequal variances to further tease apart the cognitive substrates of stochastic satisficing.

Experiment 2

To conduct a more rigorous test of the parsimony and plausibility of the stochastic satisficing heuristic, in Experiment 2, we designed a new payoff structure for the two-arm bandit, focusing on options with unequal variances in all conditions (Fig. 3A). We kept the mean and variance of the bad option constant across conditions (mean=35 and variance=10^2=100) while varied the mean and variance of the better option in a 2x2 design. Mean reward of the good option could be low (mL=57; still better than the bad option) or high (mH=72), and its variance could be independently low (vL=5^2=25) or high (vH=20^2=400). Thus, we constructed four experimental conditions all involving options with unequal variances and large or small differences in their expected values. Given this design, we simulated each of the four models, with free parameters fitted to data from experiment 1, and drove each model's predictions for choice and confidence for Experiment 2 (see SI Appendix).

Model Predictions

The most striking qualitative difference between the models was in their predictions of confidence reports when choosing the good option. All models predicted the lowest confidence for choosing the good option with the low mean and high variance ('mL-vH') (Fig. S3). Highest confidence was predicted when choosing the good option with high mean and

low variance ('mH-vL') by all models. However, SSAT model was the only one predicting similar confidence ratings for the low mean, low variance ('mL-vL') and the high mean, high variance ('mH-vH') conditions, as the probability of exceeding the satisficing threshold was the same for these two conditions. Critically, because these two conditions have different expected rewards, the 'Reward' and 'Risk' models predicted different confidence levels for these two conditions. 'Utility' model, penalizing outcomes values according to variance, predicted that higher mean and higher variance options will be chosen much less than lower mean and lower variance 'mL-vH' < 'mL-vL' > 'mH-vH' < 'mH-vL'.

Behavioural data

We examined participants' choices and confidence reports in experiment 2 in a new group of subjects (N=33) and compared them to the predictions made by the models. The probability of choosing the good option increased with the mean reward of the good option (repeated measures ANOVA, $F(1,89)=21.25$, $p=0.0001$) and decreased with its variance ($F(1,89)=15.03$, $p=0.0005$) (Fig. 3C). Confidence when choosing the bad option did not change significantly across conditions (Fig. 3C, right panel). When choosing the good option, confidence was significantly affected by variance ($F(1,89)=35$, $p<0.00001$) and mean ($F(1,89)=88$, $p<0.00001$) of the good option's rewards. However, while confidence in the 'mH-vL' was significantly higher than all other conditions ('mH-vL' vs 'mH-vH': $t(30)=4$, $p=0.0003$), and confidence in the 'mL-vH' was lower than all other conditions ('mL-vH' vs. 'mL-vL': $t(30)=4.9$, $p=0.00002$), the critical comparison of 'mL-vL' and 'mH-vH' did not show a difference ('mL-vL' vs. 'mH-vH': $t(30)=0.4$, $p=0.68$). This pattern, for both confidence reports and choices, was in line with the SSAT model predictions. Importantly, SSAT model was the only model predicting similar decision confidence in these two conditions.

To qualitatively evaluate these observations, we measured the correspondence between the pattern of confidence reports made by participants and the pattern predicted by each model through simulations. We regressed individual responses from the mean simulated pattern and obtained individual goodness of fit, R^2, for each model (see Methods). Importantly, and confirming the qualitative observations, the SSAT model gave the best prediction to the confidence reports pattern (mean±SEM: SSAT $R^2=0.68\pm0.05$, 'Reward' $R^2=0.62\pm0.06$, 'Risk' $R^2=0.63\pm0.05$, 'Utility' $R^2=0.4\pm0.04$; paired t-test: SSAT vs. 'Reward': $t(30)=2.18$; $p=0.03$; SSAT vs. 'Risk': $t(30)=2.19$; $p=0.03$) and SSAT vs. 'Utility' model ($t(30)=7.4$; $p=2*10^{-8}$).

Discussion

We set out to examine decision-making and confidence reports in uncertain value-based choices. In a two-armed bandit task played by human subjects, the probability of choosing the good option increased as a consequence of decreasing the variance of either options' outcomes. However, confidence ratings were associated with variance only when choosing

the good (higher mean) option, as items with low variance outcomes were chosen with higher decision confidence. Confidence ratings associated with choosing the bad (i.e. lower mean) option were always low and were independent of the variances of the options' outcomes. To account for these patterns, we proposed a stochastic satisficing (SSAT) model in which decisions are made by comparing the options' probability of exceeding an acceptability threshold and confidence reports scale with the chosen option's satisficing probability. To directly test a critical prediction of this model, a second experiment involving options with unequal variances and means was simulated first and then empirically performed. As predicted by the SSAT model, participants' confidence reports matched the options' probability of exceeding a threshold, and not the options' expected outcome.

We compared our model with three alternative accounts of the influence of outcome variance on choice and confidence. The simplest model assumed that choice is governed only by the expected reward, and tracked the mean outcomes of options over time [16,17]. Outcome variance may make the learning process noisier, as samples are more variable, but ultimately this model's predictions were not affected much by outcome's variance and clearly at odds with our behavioural results. The second alternative model tested the possibility that outcome variance may affect behaviour if outcome valence (positive vs negative) affects learning [1]. Increased variance in our case meant that the more uncertain option could yield both very good and very bad rewards. A risk aversive agent that learns more from negative outcomes would therefore penalize an uncertain outcome. Another way of modelling the impact of risk aversion on variance is suggested by utility theory [3,15]. Following utility theory prescription, rewards are transformed using a concave utility function. When applied to continuous outcomes with Gaussian distribution, maximizing expected utility boils down to penalizing outcomes according to their variance. An important feature of both of these instantiations of risk aversion (valance dependent learning rates and utility theory) is that the effect of variance is always in the same direction, reducing the value or utility of both good and bad options. This means that when the variance of the bad option increases, the likelihood of choosing the good option should increase. This is clearly not the case in our experimental results. Our stochastic satisficing model is unique among the models tested here by providing a mechanism by which variance effect is not symmetrical for good and bad option – when the bad option's variance increases, its value (i.e. the probability of surpassing an acceptability threshold) increases.

A number of recent studies [11,14,22,23] have formulized confidence as the probability of having made a correct choice over tracked outcome or evidence distribution. This approach builds on the line of research about the representation of evidence distribution, and suggests that confidence summarizes this probabilistic representation, estimating the probability of being correct. Probability of being correct is more readily defined in perceptual detection tasks where option outcomes are not independent (e.g. the target can be in only one of two locations but not both) and there is an objective criterion for correctness. When the two

options have continuous outcome and are governed by independent distributions, confidence was shown to reflect the difference in values between the two choice items [20,24]. However, in these studies, items' values were predefined, and importantly had no uncertainty associated with them. Our stochastic satisficing model combines these two approaches, expanding observations from value-based decisions to scenarios where outcomes are stochastic. In such scenarios, our theory-based analysis of data suggests, participants use an arbitrary criterion, the acceptability threshold, to evaluate the probability of an outcome to exceed the threshold, analogous to the evaluation of correctness probability in detection tasks. Confidence would then reflect the probability that the chosen option exceeded the "good enough" acceptability threshold. As the likelihood of exceeding the acceptability threshold increases – either by reducing the outcome variance (Experiment 1) or increasing the outcome mean (Experiment 2) – so does decision confidence.

Research in administrative decision making suggests that when faced with elaborate and complex environments, where finding the optimal solution to a problem can prove to be very costly in terms of time, money and cognitive resources, people employ heuristics when making decisions [5,8,10]. In the 1950s Simon introduced the concept of satisficing, by which decision makers settle for an option that satisfies some threshold or criterion instead of finding the optimal solution. The idea is illustrated in the contrast between 'looking for the sharpest needle in the haystack' (optimizing) and 'looking for a needle sharp enough to sew with' (satisficing) (p. 244) [10,25]. This notion of acceptability threshold has been extended to other ambiguous situations [26], for example for setting a limit (i.e. threshold) to the time and resources an organization invests in learning a new capability [10], where suboptimal solution may be balanced with preventing unnecessary cost.

We suggest that our stochastic satisficing serves a similar objective by extending the basic idea of satisficing into stochastic contexts with continuous payoff domains. While optimality prescribes choosing the option with the highest expected value, a general solution for computing this quantity (i.e., expected value), given an arbitrary distribution over the payoff, is computationally expensive. For continuous distributions, such normative solution would require computing the integral over multiplication of reward and its subjective probability. Computing the satisficing probability, however, only requires computing the area under the payoff distribution, above the acceptability threshold. Such computation is less expensive, and more receptive to heuristic estimations. In addition, stochastic satisficing may serve some other psychological and social purposes associated with decision-making. As it strives to avoid catastrophe, i.e. receiving a reward below acceptability threshold, stochastic satisficing may be useful to minimize regret, similarly to status quo bias [27,28]. Such strategy may also be useful from an accountability point of view, as choosing the option less likely to provide unacceptable payoffs can serve as a safe argument for

justifying decisions to oneself or others [29], in the spirit of the saying "nobody ever got fired for buying IBM".

Methods

Participants

88 and 33 subjects participated in the first and the second experiments, respectively, using online Amazon M-Turk. All participants provided an informed consent (experiments were approved by the local ethics committee). Participants earned a fixed monetary compensation, but also a performance-based bonus if they collected more than 10,000 points. 25 participants were excluded from analysis as their performance was at chance level (16 participants) or for using only one level for confidence reports (9 participants). Data from 96 participants (62 male aged 32±9 (mean±std); and 34 female aged 32±8) were analysed.

Experimental Procedure and Design

On each trial participants chose between two doors, each leading to a reward between 1 and 100 points (Fig. 1A). Each door had a fixed color-pattern along the task, but the positions (left vs. right) were chosen randomly. Subjects made choices by using a 12-level confidence ratings: 1-6 towards one option and 1-6 towards the other, with 6 indicating 'most certain' and 1 indicating 'most uncertain'. Following choice, subjects observed the reward of the chosen door drawn from a normal distribution $N(\mu_i,\sigma_i^2)$, where *i* was *a* or *b*, indicating one or the other door.

Experiment 1 consisted of 240 trials and included six stable blocks where the mean and variance of each option's reward remained constant. Each block lasted at least 25 trials. The transition from one block to another occurred along 10 trials during which the mean and variance associated with each door changed gradually in a linear fashion, from their current to the new levels corresponding to the upcoming block. Embedded within these six blocks, four blocks followed a 2x2 design where the mean rewards of the two options were 65 (for the good option) and 35 (for the bad option), and their variances could be independently high (H=25^2=625) or low (L=10^2=100) (Fig. 1B). This design creates four conditions: 'L-L','H-L','L-H' and 'H-H', where the first and the second letters indicated the magnitude of the variance of the good and the bad options, respectively.

Experiment 2 consisted of 160 trials and was similarly composed of blocks of fixed reward probability distributions. In all four blocks, the reward of one option always followed a Gaussian distribution with a mean of 35 and a variance of 100 (10^2). The mean of the other option could take either high (mH=72) or low (mL=57), and its variance could be either high (vH =20^2=400) or low (vL=5^2=25). This produced a 2x2 design, with the four conditions denoted by 'mL-vL', 'mL-vH', 'mH-vH', 'mL-vL'(Fig. 3A).

SSAT Model

Stochastic satisficing (SSAT) model employs a threshold heuristic. It uses the means and variances associated with the two options, to compare the probability of each option's payoff exceeding an acceptability threshold. These probabilities are compared using a softmax decision rule.

Tracking the mean of rewards is done using a temporal difference algorithm [16,17] (eq. 1). Upon receiving a reward, *R(t)*, the mean of the observed option is updated with learning rate α, whereas the unselected option's mean is forgotten over time and drifts towards the acceptability threshold, *T*:

$$\begin{cases} Q_a(t+1) = Q_a(t) + \alpha \cdot \left(R(t) - Q_a(t)\right) \\ Q_b(t+1) = Q_b(t) + \alpha \cdot \left(T - Q_b(t)\right) \end{cases} \quad [1]$$

Tracking the variance is done using a similar temporal difference rule:

$$\begin{cases} V_a(t+1) = V_a(t) + \gamma \cdot \left(\left(R(t) - Q_a(t)\right)^2 - V_a(t)\right) \\ V_b(t+1) = V_b(t) \end{cases} \quad [2]$$

Where γ is the variance learning rate. The probability of payoff being higher than the acceptability threshold, *T*, was calculated using a cumulative Gaussian distribution equation:

$$SPa(t) = \frac{1}{\sqrt{2\pi V_a(t)}} \int_T^{\infty} \exp\left(-\frac{\left(x - Q_a(t)\right)^2}{2V_a(t)}\right) dx \quad [3]$$

where SP_a indicated the probability of action *a* being satisficing. A softmax rule was used to calculate choice probabilities according to the options' satisficing probabilities (SP_a and SP_b):

$$p_a(t) = \frac{\exp\left(\beta \cdot SPa(t)\right)}{\exp\left(\beta \cdot SPa(t)\right) + \exp\left(\beta \cdot SPb(t)\right)} \quad [4]$$

where β is the rate of exploration (inverse temperature). The SSAT model, thus, has 4 free parameters: {α, γ, β, T}.

Acknowledgements

UH and BB are supported by the European Research Council (NeuroCoDec 309865). UH is also supported by the John Templeton Foundation. MK is supported by the Gatsby Charitable Foundation.

References

1. Niv, Y., Edlund, J. a, Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32,** 551–62 (2012).
2. Erev, I. & Barron, G. On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychol. Rev.* **112,** 912–931 (2005).
3. Camerer, C. F., Loewenstein, G. & Rabin, M. *Advances in behavioral economics*. (Princeton University Press, 2011).
4. Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision under RisK. *Econometrica* **47,** 263–292 (1979).
5. Tversky, A. & Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science (80-.).* **185,** 1124–1131 (1974).
6. Glimcher, P. W. Indeterminacy in brain and behavior. *Annu. Rev. Psychol.* **56,** 25–56 (2005).
7. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37,** 205–220 (2014).
8. Simon, H. a. Rational choice and the structure of the environment. *Psychol. Rev.* **63,** 129–138 (1956).
9. Winter, S. idney G. Satisficing , Selection , and The Innovating Remnant. *Q. J. Econ.* **85,** 237–261 (1971).
10. Winter, S. G. The Satisficing Principle in Capability Learning. *Strateg. Manag. J.* **21,** 981–996 (2000).
11. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90,** 499–506 (2016).
12. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329,** 1541–3 (2010).
13. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 1310–21 (2012).
14. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19,** 366–374 (2016).
15. Sargent, T. J. *Macroeconomic theory*. (1979).
16. Rescorla, R. A. & Wagner, A. R. in *Classical conditioning II: current research and theory* (eds. Black, A. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, 1972).
17. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. **1,** (MIT press, 2012).
18. Kruschke, J. K. Bayesian Estimation Supersedes the t Test. *J. Exp. Psychol. Gen.* **142,** 573–603 (2012).
19. Watanabe, S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.* **11,**

3571–3594 (2010).

20. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16,** 105–10 (2013).

21. Lebreton, M., Jorge, S., Michel, V., Thirion, B. & Pessiglione, M. An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron* **64,** 431–439 (2009).

22. Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* **11,** 55–60 (2016).

23. Meyniel, F., Schlunegger, D. & Dehaene, S. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Comput. Biol.* **11,** e1004305 (2015).

24. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18,** 1159–1167 (2015).

25. Simon, H. A. *Models of bounded rationality: Empirically grounded economic reason*. **3,** (MIT press, 1982).

26. Sanborn, A. N. & Chater, N. Bayesian Brains without Probabilities. *Trends Cogn. Sci.* **xx,** 1–11 (2016).

27. Nicolle, A., Fleming, S. M., Bach, D. R., Driver, J. & Dolan, R. J. A Regret-Induced Status Quo Bias. *J. Neurosci.* **31,** 3320–3327 (2011).

28. Samuelson, W. & Zeckhauser, R. Status quo bias in decision making. *J. Risk Uncertain.* **1,** 7–59 (1988).

29. Lerner, J. S. & Tetlock, P. E. Accounting for the effects of accountability. *Psychol. Bull.* **125,** 255–275 (1999).

Figures and Tables

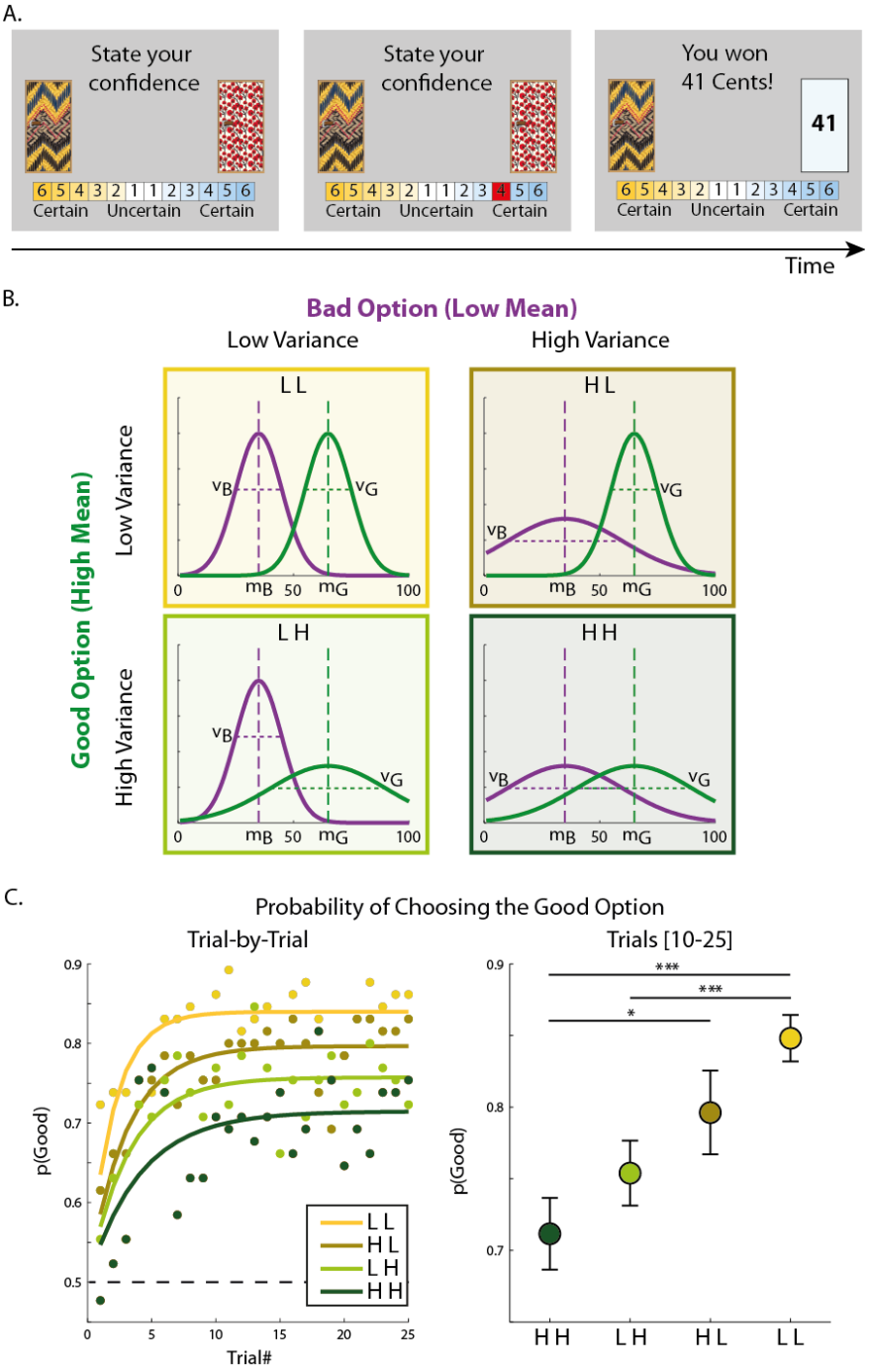

Figure 1: Design and results of experiment 1. (A) Four different experimental conditions embedded in a continuous two-armed bandit task. In each condition one door had a low (m_B, Bad option) and the other had a high (m_G, Good option) expected reward. Mean reward (m_B and m_G) were constant across conditions. The variances of the two distributions, however, changed across conditions and was either high or low, resulting in a 2x2 design (V_B (Low/High) x V_G (Low/High)). Each condition lasted between 27 to 35 consecutive trials. (B) The probability of choosing the good option (averaged over 65 participants) in each

experimental condition, across trials (left panel) and averaged across trials 10 to 25 (right panel). (* $p<0.05$, *** $p<0.0005$).

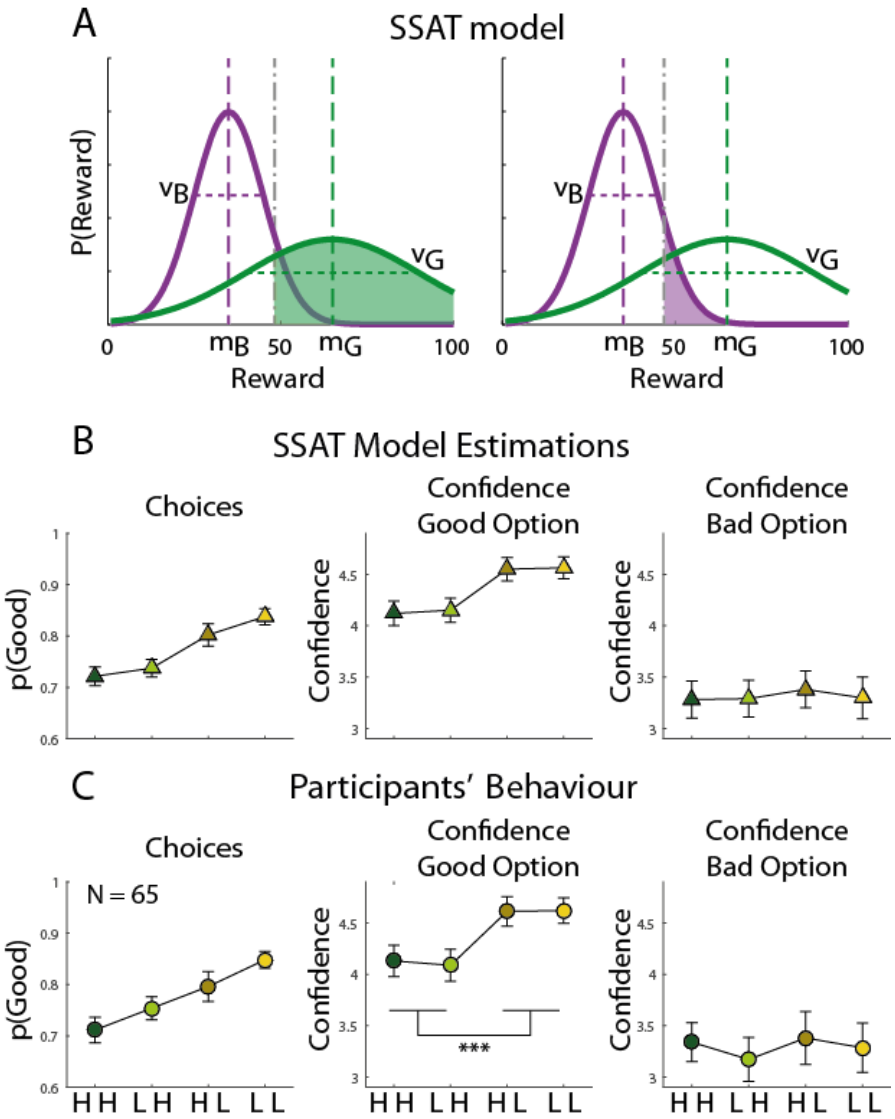

Figure 2: Model predictions and experimental results for experiment 1. (A) Stochastic satisficing model suggests that decisions are made based on the probability of each door's outcome exceeding an acceptability threshold (grey dot-dashed line). This probability (area under the curve) is higher for the door with the high mean expected reward (left) than for the door with the low mean (right). It decreases as variance increases. (B) Estimated results of the stochastic satisficing model (best fit to trial by trial choice data) under different experimental conditions, for choice (left panel), and predictions of confidence reports when choosing the good (middle panel) or bad (right panel) option. Models' estimated confidence ratings were rescaled individually to range between 1 and 6 for presentation purposes (see Methods). When choosing the good option, SSAT model predicts higher confidence when variance is low compared with high variance, even though mean expected reward remained constant. (C) Experimental results (65 subjects). Both model estimations and experimental results were averaged between trials 10 to 25 of each experimental block. When choosing the good option, confidence ratings were higher when variance of the good option was low, regardless of the variance of the bad option. (*** $p<0.0005$).

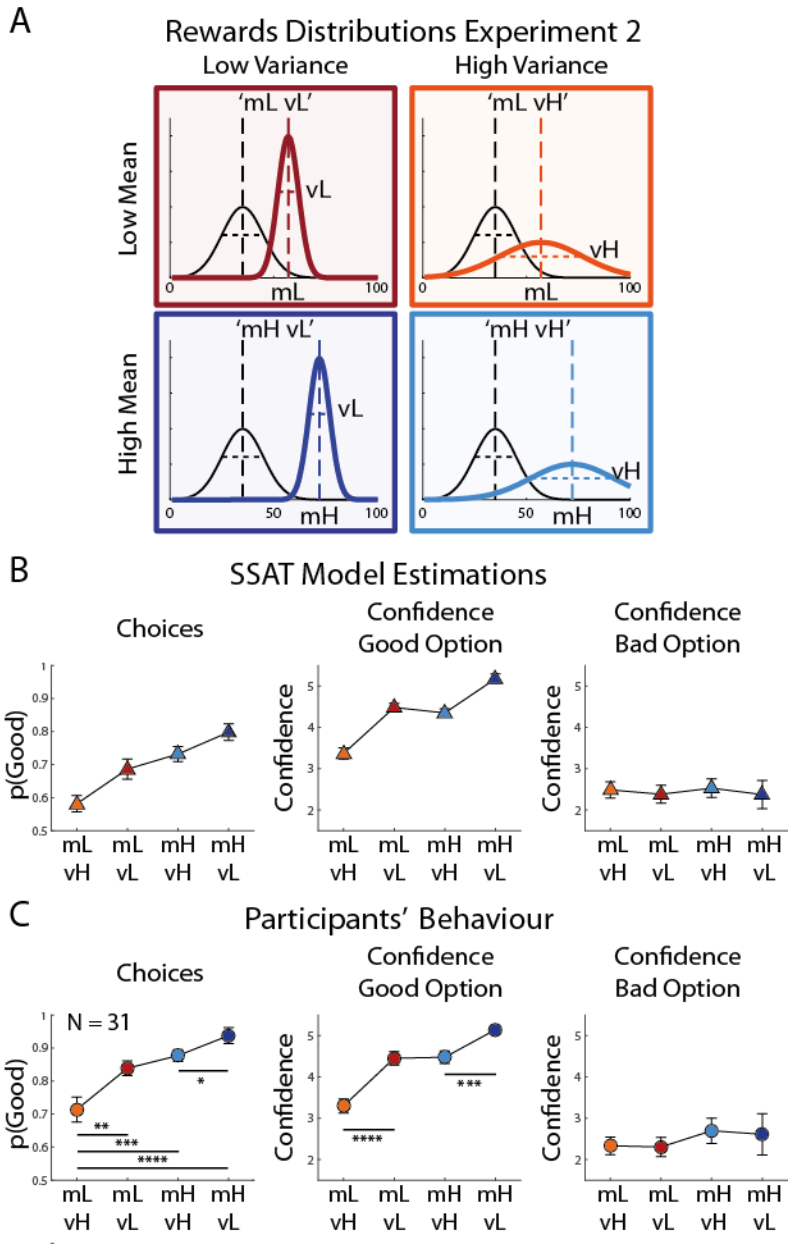

Figure 3: Model predictions and experimental results for experiment 2. (A) In experiment 2 the reward's mean and variance of the bad option (black lines) were kept constant across experimental conditions, while the mean and variance of the good option varied. Mean values could be high (mH) or low (mL), and variances could be independently high (vH) or low (vL), resulting in four experimental conditions. (B) Simulating the SSAT model (best fit to choice data from experiment 1) predicted similar confidence reports when choosing the good option in the 'mL-vL' and 'mH-vH' conditions. Models' estimated confidence ratings were averaged between trials 10 to 25, and rescaled individually to range between 1 and 6 for presentation purposes (see Methods). (C) Experimental results (33 subjects). Both choices and confidence reports were averaged between trials 10 to 25 of each experimental block. When choosing the good option (middle panel), confidence ratings were highest did not differ between the 'mL-vL' and 'mH-vH' condition (* $p<0.05$, *** $p<0.0005$).