Apparently Irrational Choice as Optimal Sequential Decision Making

Haiyang Chen,[1] **Hyung Jin Chang,** [1] **Andrew Howes** [1,2]
[1] School of Computer Science, University of Birmingham
[2] Department of Communications and Networking, Aalto University
{hxc797, h.j.chang, howesa}@bham.ac.uk

Abstract

In this paper, we propose a normative approach to modeling apparently human irrational decision making (cognitive biases) that makes use of inherently rational computational mechanisms. We view preferential choice tasks as sequential decision making problems and formulate them as Partially Observable Markov Decision Processes (POMDPs). The resulting sequential decision model learns what information to gather about which options, whether to calculate option values or make comparisons between options and when to make a choice. We apply the model to choice problems where context is known to influence human choice, an effect that has been taken as evidence that human cognition is irrational. Our results show that the new model approximates a bounded optimal cognitive policy and makes quantitative predictions that correspond well to evidence about human choice. Furthermore, the model uses context to help infer which option has a maximum expected value while taking into account computational cost and cognitive limits. In addition, it predicts when, and explains why, people stop evidence accumulation and make a decision. We argue that the model provides evidence that apparent human irrationalities are emergent consequences of processes that prefer higher value (rational) policies.

Introduction

Recent efforts toward a computational understanding of the human mind have been invigorated by advances in Artificial Intelligence (AI) (Lewis, Howes, and Singh 2014; Cichy and Kaiser 2019; Lieder and Griffiths 2019; Gershman, Horvitz, and Tenenbaum 2015). As machine learning has progressed, reinforcement and deep learning algorithms have generated systems that attained human- and superhuman-level performance in a number of domains and it is believed by many researchers that modern AI not only has the capacity to equal human performance but also to help inform deeper understandings of human cognition (Hassabis et al. 2017; Lake et al. 2017). In other words, building computational models of human cognition, informed by modern machine learning, offers a potential way to advance our understanding of cognitive processes (McClelland 2009; Fontanesi et al. 2019; Lieder and Griffiths 2017; Lieder, Griffiths, and Goodman 2012; Milli, Lieder, and Griffiths 2017).

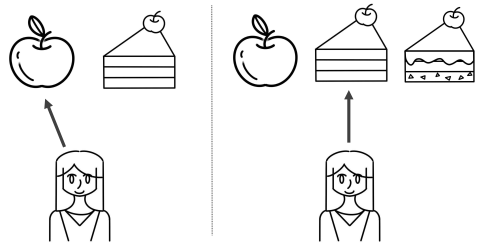

Figure 1: An example of a context-of-choice bias. If a person chooses an apple over a cake on the grounds of health, but then chooses the same cake when the choice is between an apple, the cake and another cake with extra sugar, then the clearly inferior (on health grounds) "cake with extra sugar" has influenced the choice between two superior alternatives.

However, the fact that humans appear to exhibit decision-making biases seems to pose a severe challenge to this contention. There are a variety of apparent human biases (Kahneman 2016) suggesting that people are not good expected value maximizers and that they are subject to irrationalities of choice that are counter to their own self interest. In this paper, we focus on context effects which show that human preference between two options changes when a third (dominated) alternative is introduced, e.g. as shown in Figure 1. As the dominated choice is irrelevant to the choice between the other two options, it should have no effect on their valuation, nor on the choice. This effect has been taken by some as evidence that human cognition is irrational since it appears to violate the normative principles of independence (Usher and McClelland 2001). A number of computational models reproduce these effects. They explain context effects with algorithms that concern what evidence to gather and how that evidence is accumulated (Stewart, Chater, and Brown 2006; Busemeyer et al. 2019; Wollschlaeger and Diederich 2020). While they offer accounts of how the dynamics of the decision making unfold they do not explain why cognitive processing would lead to apparently irrational responses.

However, recent research has begun to show that people may exhibit rationality more often than supposed (Lewis, Howes, and Singh 2014; Chen 2015; Chen et al. 2017; Lieder and Griffiths 2019; Todd and Gigerenzer 2012; Fra-

zier and Angela 2008; Tsetsos et al. 2016; Howes et al. 2016). In response, we present a normative decision-making model for contextual choice tasks based on POMDPs, which provides a unifying framework for modelling various fundamental cognitive components of human decision making required to explain contextual preference reversal. Our work is inspired by demonstration (Howes et al. 2016) that apparent irrationalities of choice can emerge from rational processing. Our approach is an application of computational rationality (Lewis, Howes, and Singh 2014; Griffiths, Lieder, and Goodman 2015; Howes, Lewis, and Vera 2009) to the problem of human decision making. It extends (Howes et al. 2016) by modeling contextual choice tasks as sequential decision problems and formulating them as POMDPs. Previous work by (Dayan and Daw 2008; Rao 2010; Frazier and Angela 2008; Howes et al. 2018) and others has established the value of POMDPs and related formalisms for modeling humans.

In our work, a Reinforcement Learning (RL) agent, designed to solve a POMDP, acquires a sequential decision policy that chooses what information to gather about which options, calculates option values, and makes comparisons between options as the unfolding task demands. The agent is trained and tested on sampled choices between three gambles, each expressed as a probability and a value. It learns the relative value of (1) noisy calculation of option values (e.g. by multiplication of a probability by a value), (2) noisy comparisons (e.g. comparing two probabilities to see which option is riskier), and (3) acting (making a choice). The agent is not pre-programmed to gather all information but learns to gather only that information that helps it maximize utility. We contrast the policies acquired by this agent to other simpler agents and show that the human-inspired agent performs better (achieves higher cumulative reward) than an agent that makes independent assessments of each option value, replicating the results of (Howes et al. 2016) but in the POMDP setting.

Our contributions are as follows:

- A computationally rational model of contextual choice formulated as a POMDP. The model shows that preference reversals are a consequence of rational policies that prefer higher value policies. To avoid confusion, it is also important to say that the model is not a model of human learning processes. It is a model of the emergent sequential decision policy.
- Novel predictions concerning optimal sequential information gathering in contextual choice tasks. In particular, our model shows how the ratio of option comparisons and expected value calculations is influenced by the level of uncertainty in the observation function.
- An extension to the analysis of (Howes et al. 2016) that accounts for the impact of sequential information gathering costs on contextual choice.
- Advancing a general understanding of how rationality, uncertainty, and apparent biases are connected. These connections are critical to the future of AI systems that work with people.

Background

The Effect of Choice Context on Humans

As we have said, the human behaviours that have influenced this paper are those exhibited in decision-making tasks in which people appear biased by seemingly irrelevant context. Here we look in more detail at these tasks and their effects. Three of the most well known contextual decision task are the attraction, compromise and similarity tasks. These are illustrated in Figure 2a, b, c. For the attraction type task, there are two best options (the Target and the Competitor) with the similar expected value. Each option is best on one dimension but not the other. One of these two options (the Target) dominates a third option, called the decoy, on both dimensions. It is difficult to choose between the two best options since each option dominates the other on one of the attributes. Experiments studying these three tasks have been reported by many authors. Consider the results of an experiment in which participants were asked to make decisions about criminal suspects (Trueblood 2012). Participants were presented with a sequence of tasks each consisting of three suspects and were asked to decide which suspect was most likely to have committed a crime. There were two types of evidence, of varying strength, about each of the three suspects, such that the suspects had likelihoods of criminality in patterns identical to the three patterns presented in Figure 2d. These three patterns were used as the materials in the three conditions of the experiment.

In the attraction condition of the experiment, there were two equally likely criminal suspects and a decoy suspect who was less likely than the other two (Figure 2a). The experimental results showed that the Target suspect who dominates the decoy was chosen more frequently than the Competitor suspect. In the compromise condition of the experiment (Figure 2b) the findings showed that the suspect who is in-between the Competitor and Decoy is chosen more frequently than the Competitor. In the similarity condition (Figure 2c), the results showed that the suspect who is very similar to the decoy is chosen less frequently than the Competitor.

These effects have contributed to shaping a number of cognitive theories of human decision making (Usher and McClelland 2001; Roe, Busemeyer, and Townsend 2001; Howes et al. 2016; Busemeyer et al. 2019; Noguchi and Stewart 2018; Ronayne and Brown 2017; Spektor et al. 2019; Wollschlaeger and Diederich 2020). In some, though not all of these theories, human behaviour has been assumed to be biased because the irrelevant context (the decoy option) has consequences for the choice between the other two options((Tversky and Simonson 1993), p. 1188). The most commonly used operationalization of irrationality among decision researchers has been based on violations of value maximization. Preferring a dominated option or expressing different preferences depending on the framing of options demonstrates irrational decisions. The significance of any irrationality, if that is what they are, cannot be understated given the potential for catastrophic real world consequences. However, the conclusion that choice under uncertainty provides evidence of irrationality may be incorrect (Howes et al.

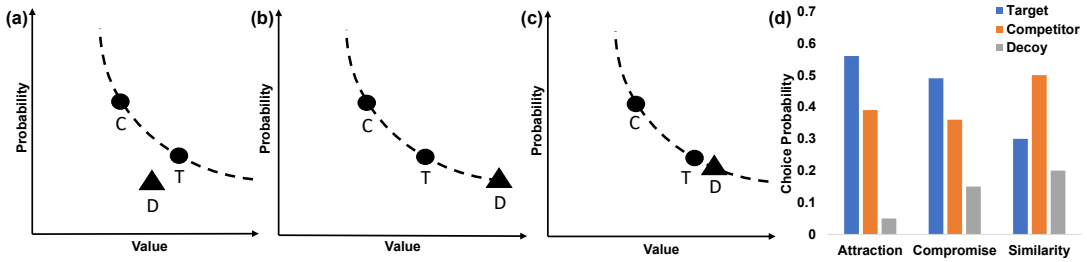

Figure 2: (a)(b)(c) An illustration of the options in three types of contextual choice task – called the attraction (a), compromise (b) and similarity (c) tasks. The Target T and Competitor C are two options and have equal expected value (the dotted line). Option D is a decoy designed for comparison with the Target T. In the attraction task (a), T dominates D. In the compromise task (b), T is a compromise between D and T. In the similarity task (c), D has similar expected value to T. (d) Proportion of choices of each of the three options (Target, Competitor and Decoy) in each of the three contextual choice tasks (Attraction, Compromise and Similarity). The Target is preferred in the Attraction and Compromise tasks and the Competitor is preferred in the Similarity task. Data are reproduced from (Trueblood 2012).

2016; Tsetsos et al. 2016). Substantive analysis of the value of comparing options has shown that they are in fact informative and are required, under conditions of uncertainty, for reward maximization (Howes et al. 2016). The substantive structure of these analyses has informed the design of the agent that we present below. The key idea that is borrowed from human behaviour is the use of option comparison to inform decision making under uncertainty. Comparison was extensively explored by Stewart (Stewart, Chater, and Brown 2006; Vlaev et al. 2011) who has documented extensive of its use in a range of human decision making tasks. For example, there is eye tracking evidence (Noguchi and Stewart 2014) that people tend to make more eye movements that switch between options than eye movements that gather all of the evidence about a single option; evidence which is consistent with the use of comparisons.

Human Decision Making as a POMDP

POMDPs provide a mathematical framework for sequential decision processes (Kaelbling, Littman, and Cassandra 1998). POMDPs have previously been used for modelling and explaining various aspects of human decision making (Daw, Courville, and Touretzky 2006; Dayan and Daw 2008; Rao 2010). An early contribution was (Daw, Courville, and Touretzky 2006)'s model of the dopamine system which incorporated semi-Markov dynamics and partial observability. (Rao 2010) proposed a model of neural information processing based on POMDPs and tested this model on perceptual tasks such as the random dot motion task. Further work in perceptual decision making, has used the POMDP framing to explore model confidence (Khalvati and Rao 2015) and understand the role of priors (Huang et al. 2012). POMDPs have also been used to model social decision making (Khalvati et al. 2016) and Theory of Mind (Baker, Saxe, and Tenenbaum 2011; Baker et al. 2017; Rabinowitz et al. 2018). More recently, meta-level Markov decision processes (meta-MDP), a closely related framework, have been used for modelling higher level decision making (Griffiths et al. 2019). The Meta-MDP model is similar to the belief-MDP version of the POMDP, but replacing physical actions with cognitive operations. Meta-MDPs have been used to model strategy selection and heuristics in decision making (Lieder, Krueger, and Griffiths 2017) and attention allocation in perception (Callaway, Antonio, and Tom 2020).

As we have said, contextual preference reversals have influenced a number of models of human decision making (Usher and McClelland 2001; Roe, Busemeyer, and Townsend 2001; Frazier and Angela 2008; Trueblood, Brown, and Heathcote 2014; Ronayne and Brown 2017; Noguchi and Stewart 2018; Busemeyer et al. 2019). Many of these models have focused on neurally plausible sequential processing, capturing the fact that decision making usually requires accumulation of evidence and integration of information across time (Tsetsos et al. 2016). Other models have focused on the way that people solve this problem by sampling comparisons between option attributes and thereby impose a rank order on options (Noguchi and Stewart 2018). However, none to our knowledge, have shown that preference reversals are an emergent consequence of an RL solution to a POMDP.

Contextual Choice as a POMDP

Unlike existing models for the contextual choice task, we present a normative decision-making model based on POMDPs, which provides a unifying framework for modelling various cognitive components of human decision making including noisy evidence accumulation, reward maximization, costs and rewards of actions, uncertainty evaluation, etc.

We view contextual choice tasks as sequential decision making problems and formulate them as POMDPs that include comparison actions to assess choice option values. Given this formulation, we use a deep reinforcement learning model to discover an approximately optimal choice policy and demonstrate its capacity to simultaneously maximize reward and model the human decision making process in contextual choice tasks. A crucial property of the model is that gathering information is costly, so that more information costs more but also increases the probability of a better, more rewarding, choice.

We start with a standard definition of a POMDP as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\mathcal{O}$ is the observation space. At each time step t the agent is in the latent state $s_t \in \mathcal{S}$, which is not directly observable to the agent. When the agent executes an action $a_t \in \mathcal{A}$, the state of the process changes stochastically according to the transition distribution, $s_t \sim T(s_{t+1}|s_t, a_t)$. Then, to gather information about the state, the agent makes a partial observation $o_{t+1} \in \mathcal{O}$ according to the distribution $o_{t+1} \sim \mathcal{Z}(o_{t+1}|s_{t+1}, a_t)$. The agent receives a reward $r_t \in \mathcal{R}$ according to the distribution $r_t \sim \mathcal{R}(s_t, a_t)$ after performing an action a_t in a particular state s_t. The agent must rely on its observations to inform action selection since the hidden states are not directly observable. In each time step t, the agent acts according to its policy $\pi(a_t|h_t)$ which returns the probability of executing action a_t, and where $h_t = (o_0, a_0, o_1, a_1, \cdots o_{t-1}, a_{t-1})$ are the histories of observations-actions pairs. The goal of the agent is to learn an optimal policy π^* that maximizes the expected cumulative rewards, $\pi^* = \underset{\pi}{argmax}\mathbb{E}\left[\sum_{t=1}^{t=T} \gamma^{t-1} r_t\right]$, where $0 < \gamma < 1$ is the discount factor.

Each choice task has 3 options (X, Y, Z) which were represented with two attributes: a randomly sampled probability p and a randomly sampled value v. We assumed that probabilities p were sampled from a $\beta-$distribution and values v were sampled from a $t-$distribution. These distributions represented the ecological distributions experienced by participants in the human behaviour experiments reported by (Wedell 1991). We view contextual choice tasks as sequential decision making problems and formulate them as POMDPs as follows.

The state space $\mathcal{S}$ for each task was generated from a sampled choice task. More formally a state was $\{(p_X, v_X), (p_Y, v_Y), (p_Z, v_Z)\}$. The agent selected actions from a set $\mathcal{A}$ which included 6 comparison actions (e.g. compute the comparison relation for p_X and p_Y), 3 calculation actions (e.g. compute the expected value for X given p_X and v_X) and the 3 choice actions (choose X, choose Y, choose Z). The reward for comparison and calculation actions was negative c. The reward for a choice action was 1 if the agent chose the option with maximum expected value, otherwise, it was -1. There was therefore a trade-off between the cost of information gathering and choice accuracy. More information cost more but was more likely to lead to a better response and therefore a higher reward. As a consequence of the selected action, the subsequent observation $o_{t+1} \in \mathcal{O}$ was of computing the most recent comparison or calculation with noise. Following (Howes et al. 2016) each observation of a comparison had 4 possible outcomes, which indicated that the relation was unknown, greater-than, equal or less-than. The function f represents this pairwise order relation between the two values or two probabilities of two gambles.

The probability of comparison error $P(error_f)$ was the probability that the relations were sampled uniformly random from the comparison set $O = \{>, \equiv, <\}$. The observation of a calculation was computed using:

$$E_i = p_i^{\alpha} \times v_i + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2_{calc}) \tag{1}$$

where the probability p was weighted by an exponential parameter α. The purpose of using parameter α was to model **subjective probability** following (Savage 1972), which is used extensively in econometrics because it is mathematically well behaved.

The "evidence" state is the history of the partial and noisy observation of the latent state. The set of possible observations in the history $\mathcal{O}_h$ is the set of noisy encodings of partial orderings and calculations:

$$\begin{aligned}\mathcal{O}_h = \{&f(p_X, p_Y), f(p_X, p_Z), f(p_Y, p_Z), f(v_X, v_Y),\\ &f(v_X, v_Z), f(v_Y, v_Z), E_X, E_Y, E_Z\}\end{aligned} \tag{2}$$

It is intractable to compute a policy to solve the defined POMDP, but it is possible to approximate the optimum through learning (Wang et al. 2018; Igl et al. 2018). We solve the POMDP by casting it as a Markov Decision Process (MDP) whose state space is the history of observation o_h. We used various deep RL methods to find the approximately optimal policy for the POMDP, e.g. ACER (Wang et al. 2016). To guarantee the generality, we used the standard RL methods to find the solution. The results show that the direction of the effects is not sensitive to the training parameter values. For all reported experiments, we built the environments within OpenAI Gym (Brockman et al. 2016) and used the Baselines[1] implementation of the deep RL algorithms.

Results

In order to test the model, we designed three different agents: The **integrated agent** could use both calculation and comparison selectively. States represented the results of calculation and comparison actions. The model could learn which observations are useful. Not every observation needed to be made. There was no explicit integration of comparison and calculation. Instead, the comparison and calculation observations accumulated in the history and choice action values were conditional on these histories. The **comparison-only agent** was the same as the integrated agent but could only use comparison actions, and states only represented the comparison information. The **calculation-only agent** was the same as the integrated agent but could only use calculation actions and, states only represented the calculation information. The important difference between the three models was the availability of the different kinds of observation. All three agents learnt approximately optimal policies from experience given the bounds imposed by these difference observation capacities.

In what follows we first show that our new reinforcement learning model replicates previous findings (Howes et al. 2016) and then show that it makes new predictions derived from the sequential nature of the model.

Is it beneficial to compare options? In order to answer this question, following (Howes et al. 2016), we first fitted the distributions of the environment to those used in a prominent human experiment (Wedell 1991). The probabilities p are $\beta-$distributed ($a = 1, b = 1$) and the values v are $t-$distributed ($location = 19.60, scale =$

[1] https://github.com/openai/baselines

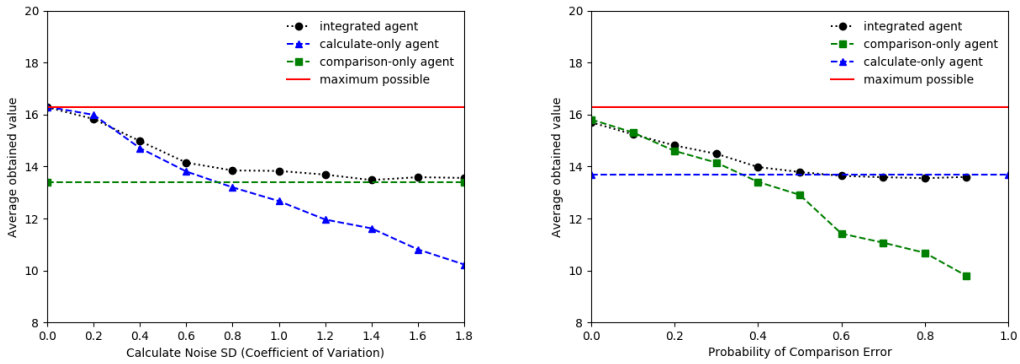

Figure 3: The mean expected value obtained by agents with different levels of noise: the coefficient of variation for the calculation noise (left panel) and the probability of comparison error (right panel). In the left panel, the comparison noise is fixed at $P_{error} = 0.3$. In the right panel, the calculation noise is fixed at $\delta_{calc} = 30$, that the coefficient of variation is 0.3. Results for 3 types of agent are presented in each panel: the comparison-only agent (green-dashed line), calculation-only agent (blue-dashed line) and integrated agent (black-dotted line). This Figure shows that the model replicates the results of Figure 3 in (Howes et al. 2016).

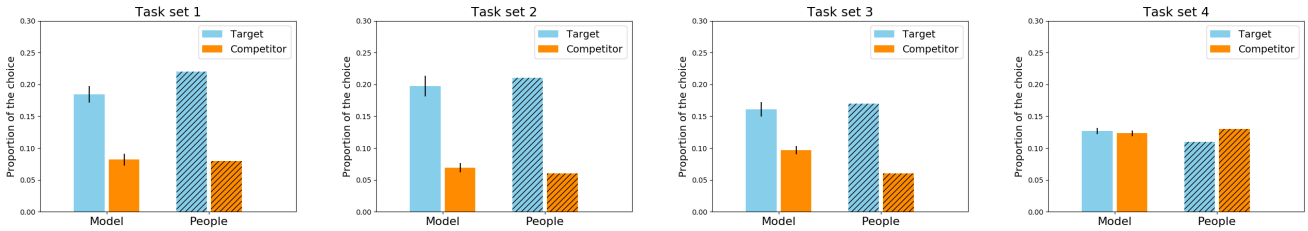

Figure 4: The integrated agent exhibits the attraction effect. A sample of agents was tested on each of four variants of the attraction effect task (in which the decoy is in slightly different positions). People and agent exhibit more target choices than Competitor choices in task sets 1, 2, and 3. As expected, neither the integrated agent, nor people, exhibit the effect in task set 4 where the decoy was not dominated by only one of the options and was therefore in a neutral position. Task 4 thereby acts as a control. The human data is from (Wedell 1991). The error bars indicate confidence intervals (95%) of the predictions made by the agent. This Figure shows that the new model replicates Figure 8 in (Howes et al. 2016).

$5, degrees\ of\ freedom = 100$). For all the experiments below, we used the same distributions. Reported results are averaged over 10 runs, each with a different seed, after training on 3 million samples. All the details on setup and learning curves can be found in the supplementary material.

All agents were tested with different levels of observation noise and the resulting performance is shown in Figure 3. The maximum expected value that could be achieved by any agent was 16.29 (horizontal upper bound in Figure 3), which was calculated by averaging the maximum expected value of 3 options across 1 million choice sets sampled from the above distributions.

In Figure 3 it can be seen that the integrated agent, using both calculation and comparison observations, can approximate the optimal policy when actions could be conducted without noise. Also, calculation-based and comparison-based agents are able to perform close to the optimum when there is no noise. However, the noise has a negative effect on the performance of all types of agent. The average obtained value of choices decreases as noise increases.

Figure 3 also shows that the integrated agent combines the strengths of both noisy comparison and noisy calculation to make better decisions than the other agents in all noise conditions. The average expected value of the choices made by the integrated agent is greater than the other agents. In other words, the human-like integrated agent performs better in accumulating reward than the agent that makes independent assessments of each option value. The results suggest that when there is observation uncertainty, both humans and artificial agents will gain higher reward if they compare options, rather than merely evaluate each option independently.

Does the integrated agent predict human performance? To determine whether the integrated agent (the agent that uses both comparison and calculation) predicts human performance, we measured its behaviour on the attraction, compromise and similarity tasks. The human behaviour on these tasks is shown in Figure 2d. We used one fixed setting of the agent policy and parameter values.

Agents were trained on tasks which were same as in the last section. After 3 million training samples, the agent con-

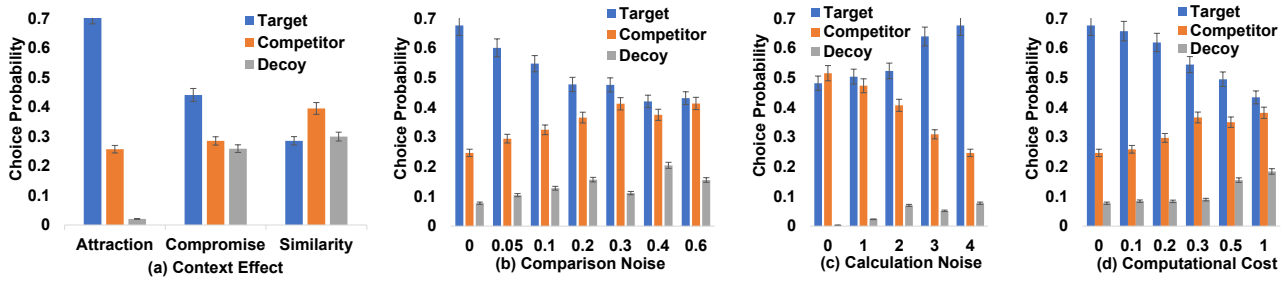

Figure 5: (a) The behaviour of the integrated agent for 3 types of context effect: attraction, compromise and similarity. (b)(c)(d) The effect of noise and computational cost on the contextual choice effect exhibited by the integrated agent. (b) Increased comparison noise reduces the effect size, (c) increased calculation noise increases the effect size, and (d) increased computational cost reduces the effect size. Error bars indicate (95%) confidence intervals.

verged and demonstrated stable performance. The agent was repeatedly trained with adjusted values of the comparison noise, calculation noise, probability weighting parameter, cost of comparisons and calculation cost until the qualitative effects fitted the human performance ((Trueblood 2013) Figure 2d).

The fitted parameter values were: calculation noise $\sigma_{calc} = 4$, comparison error $P(error_f) = 0.1$, probability weighting parameters $\alpha = 1$, the perceived cost of comparison $C_{comparison} = -0.01$ and the calculation cost $C_{calc} = -0.1$. We do not claim to have achieved the best possible fit, nor a better fit than other models. The point of the fit was to show that the qualitative effects exhibited by humans was within the space of behaviours generated by the agent.

The results are averaged over 10 runs each with a different seed and shown in Figure 5a. It shows that the agent generates the three context effects using one learnt policy and one fixed set of parameter values. Comparison of Figure 5a to Figure 2d shows that all of the qualitative effects are predicted.

To further test the agent we fitted it to human performance on variations of the attraction effect task reported in (Wedell 1991). The results in Figure 4 show that for both agents and people, the Target is selected more often than the Competitor in three of the task sets (1, 2, and 3). In contrast, and as expected, the Target and Competitor are selected equally often in the 4th task set by both agents and people. The decoy was positioned in a neutral position in task set 4 and does not therefore have an effect on the target choice rate. The fitted values were: calculation noise $\sigma_{calc} = 0.50$, comparison error $P(error_f) = 0.1$, probability weighting parameters $\alpha = 1.5$, the perceive cost of comparison features $C_{comparison} = -0.01$ and the calculation cost $C_{calc} = -0.1$.

What is the effect of noise and computation cost on the contextual choice effect? The results in Figure 5b, c, d show that: (b) the attraction effect is weaker when the agent's accuracy of comparison is diminished with noise, (c) The effect is stronger when calculation noise is higher, (d) The effect size decreases as computational cost increases. While, there is no human data that directly tests the effect of noise. There are a number of studies reporting that the rate of context effect diminishes as time pressure increases (Pettibone 2012; Trueblood, Brown, and Heathcote 2014). As shown in Figure 5b, c, d, the effects of time pressure on humans is consistent with the effect of increased noise in the model.

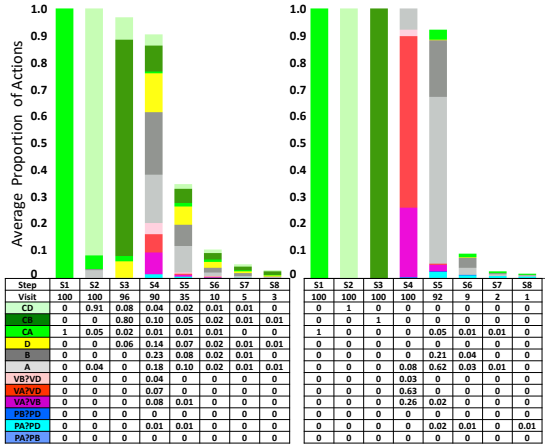

Figure 6: The left panel shows the average proportion of each action type taken by the model on each step when given randomly sampled tasks. The right panel shows the average proportion of each action type when given attraction effect preference reversal tasks. Actions that calculate the expected value of A, B or D are in green; actions that compare probabilities are shown in blue; actions that compare values are shown in red; actions that choose A, B or D are shown in white, grey and yellow respectively.

It is also worth noting that the effect sizes in Figure 5b, c, d range from 0 to over 0.43. Given that the average human attraction effect size in (Trueblood 2012) is 0.17 and in (Wedell 1991) is 0.14 this suggests that the model has sufficient scope to fit individual participants.

The effect of noise on the number of comparisons and calculation actions taken is shown in the supplementary material. Increases in comparison noise leads to a selective reduction in the use of comparison and a selective increase in the use of calculation. Conversely, increases in calculation noise leads to a selective decrease in the use of calculation and an increase in the use of comparison. Increase in the cost

of information gathering actions (comparison and expected value) reduces contextual effects on choice as less information is gathered.

How does context affect decision sequence? A novel contribution of the model is that, by virtue of the sequential decision process, it predicts how action sequences should vary with task type. Figure 6 contrast the model's action sequences on random tasks (left panel) and its action sequences on preference reversal tasks (right panel). Comparing the left and right panels, we can see that the model tends to use calculations of expected value in the first three steps regardless of task type. Despite this initial similarity, the fourth action is quite different for the two task types. Here, on average, for random tasks the model tends to pick one of the options. In contrast, for preference reversal tasks, the model tends to compare values and subsequently, shows a marked preference for the high probability option (option A). This preference is not visible in the random task action sequences. This, approximately bounded optimal, prediction conflicts with authors who have argued that people prefer comparisons to calculation of expected value (Stewart, Chater, and Brown 2006; Vlaev et al. 2011; Ronayne and Brown 2017; Noguchi and Stewart 2018).

Discussion

We have proposed a novel explanation for how apparently irrational choice might emerge as a consequence of optimal sequential decision making under uncertainty. While ours is not the first work to demonstrate the rationality of preference reversal phenomena (Howes et al. 2016), nor the first work to use POMDPs to model humans (Daw, Courville, and Touretzky 2006; Rao 2010), it is the first to formulate the contextual choice problem as a POMDP and demonstrate that a reinforcement learning agent that uses *comparison* observations generates higher reward than an agent that only makes independent assessments of value. These comparison actions, when deployed by people, have been thought by many to lead to violations of the independence axioms and they have been shown to underpin preference reversals in humans (Noguchi and Stewart 2014). But, as has previously been pointed out (Howes, Lewis, and Vera 2009), our seemingly paradoxical results make sense when it is appreciated that the comparison of options reduces the uncertainty of option values.

A different RL model of preferences reversals is reported by (Spektor et al. 2019). They explain context effects in a decisions-from-experience setting, where attribute values are are not explicitly stated but have to be learned over many trials. Our model in contrast, is based on decisions from description, where all attribute values are fully described. Unlike our model, their model does not acquire an explicit representation of different attributes and does not make attribute-based comparisons. Instead, it models a dynamic learning process during which the feedback on similar options is compared.

By extending (Howes et al. 2016) we have demonstrated that the same pattern of behaviours that are thought to be irrational in humans, emerge from a process that attempts to maximize the cumulative reward of action. Our results also show that comparison actions are increasingly *preferred* by the agent as observation noise increases. In addition, we have shown that higher information gathering costs can diminish the use of comparisons and reduce the preference reversal rate, thereby extending previous analysis to account for the economics of information gathering in contextual choice tasks. In contrast to previous models, where comparisons have been assumed, our model uses them preferentially depending on the structure of the task.

Our model assumes that observations can be subject to noise and this assumption is worth further discussion given how easy it seems for people to make comparisons. We make three observations. First, noise helps explain the fact that people make more errors when under time pressure (Pettibone 2012). These errors include choosing the distractor which is strictly dominated by one of the other choices. Comparison noise is one explanation for this error: If people select the distractor then the cannot have made correct comparisons. Second, the qualitative effects of context on preference reversal are not changed by the value of the comparison noise. All of the context effects reported in the paper are also predicted by a model without comparison noise, as shown in Figure 5b. Third, the level of comparison noise in our fitted model is so low that it results in a decoy selection rate of about 2%. A decoy selection is the only type of error in the task. This rate exactly corresponds to the human rate.

The approach that we have taken in this paper is an example of a broader class of analysis known as Computational Rationality (Lewis, Howes, and Singh 2014; Lieder and Griffiths 2019; Howes, Lewis, and Vera 2009). This approach starts from the assumption that people are approximately rational given the bounds imposed by the computation required for cognition. It then seeks to discover the computational limits that give rise to boundedly optimal (Russell and Subramanian 1994) but apparently irrational behaviours. This aim demands that the analyst derive bounded optimal policies for well-formed decision problems. Our results suggest an answer to the paradox of why it is worth motivating machine learning algorithms with apparently biased human decision making. While the behaviour appears biased, the underlying processes and heuristics (e.g. the use of option comparison) lead to gains in efficiency and therefore reward. Important directions for future research suggest that human irrationalities may offer a productive source of inspiration for improving the design of AI architectures and machine learning methods. As others have shown (Simsek, Algorta, and Kothiyal 2016) comparison observations are a particularly important avenue for exploration.

What is more, our results contribute to a growing body of work calling into question the long list of apparent irrationalities reported in the Economic literature. More may be amenable to POMDP, Meta-MDP, or MDP explanations and turn out to be rational adaptations to environmental and cognitive limits.

In conclusion, framing contextual choice problems as POMDPs reveals that apparently irrational choice reversals in behaviour are demonstrably rational under bounds imposed by uncertainty in the observation function.

Acknowledgments

We would like to thank Xiuli Chen and Richard L. Lewis for helping to develop these ideas.

References

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4): 1–10.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* .

Busemeyer, J. R.; Gluth, S.; Rieskamp, J.; and Turner, B. M. 2019. Cognitive and neural bases of Multi-Attribute, Multi-Alternative, Value-based decisions. *Trends in cognitive sciences* .

Callaway, F.; Antonio, R.; and Tom, G. 2020. Fixation patterns in simple choice are consistent with optimal use of cognitive resources. *PsyArXiv preprint PsyArXiv: https://doi.org/10.31234/osf.io/57v6k* .

Chen, X. 2015. *An optimal control approach to testing theories of human information processing constraints*. Ph.D. thesis, University of Birmingham.

Chen, X.; Starke, S. D.; Baber, C.; and Howes, A. 2017. A cognitive model of how people make decisions through interaction with visual displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1205–1216. ACM.

Cichy, R. M.; and Kaiser, D. 2019. Deep neural networks as scientific models. *Trends in cognitive sciences* 23(4): 305–317.

Daw, N. D.; Courville, A. C.; and Touretzky, D. S. 2006. Representation and timing in theories of the dopamine system. *Neural computation* 18(7): 1637–1677.

Dayan, P.; and Daw, N. D. 2008. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* 8(4): 429–453.

Fontanesi, L.; Gluth, S.; Spektor, M. S.; and Rieskamp, J. 2019. A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review* 26(4): 1099–1121.

Frazier, P.; and Angela, J. Y. 2008. Sequential hypothesis testing under stochastic deadlines. In *Advances in neural information processing systems*, 465–472.

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

Griffiths, T. L.; Callaway, F.; Chang, M. B.; Grant, E.; Krueger, P. M.; and Lieder, F. 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences* 29: 24–30.

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

Hassabis, D.; Kumaran, D.; Summerfield, C.; and Botvinick, M. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95(2): 245–258.

Howes, A.; Chen, X.; Acharya, A.; and Lewis, R. L. 2018. Interaction as an emergent property of a partially observable markov decision process. *Computational interaction design* 287–310.

Howes, A.; Lewis, R. L.; and Vera, A. 2009. Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological review* 116(4): 717.

Howes, A.; Warren, P. A.; Farmer, G.; El-Deredy, W.; and Lewis, R. L. 2016. Why contextual preference reversals maximize expected value. *Psychological review* 123(4): 368.

Huang, Y.; Hanks, T.; Shadlen, M.; Friesen, A. L.; and Rao, R. P. 2012. How prior probability influences decision making: A unifying probabilistic model. In *Advances in neural information processing systems*, 1268–1276.

Igl, M.; Zintgraf, L.; Le, T. A.; Wood, F.; and Whiteson, S. 2018. Deep Variational Reinforcement Learning for POMDPs. In *International Conference on Machine Learning*, 2122–2131.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101(1-2): 99–134.

Kahneman, D. 2016. 36 Heuristics and Biases. *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about their Most Important Contributions* 171.

Khalvati, K.; Park, S. A.; Dreher, J.-C.; and Rao, R. P. 2016. A probabilistic model of social decision making based on reward maximization. In *Advances in Neural Information Processing Systems*, 2901–2909.

Khalvati, K.; and Rao, R. P. 2015. A Bayesian framework for modeling confidence in perceptual decision making. In *Advances in neural information processing systems*, 2413–2421.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40.

Lewis, R. L.; Howes, A.; and Singh, S. 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science* 6(2): 279–311.

Lieder, F.; Griffiths, T.; and Goodman, N. 2012. Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems*, 2690–2798.

Lieder, F.; and Griffiths, T. L. 2017. Strategy selection as rational metareasoning. *Psychological Review* 124(6): 762.

Lieder, F.; and Griffiths, T. L. 2019. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 1–85.

Lieder, F.; Krueger, P. M.; and Griffiths, T. 2017. An automatic method for discovering rational heuristics for risky choice. In *CogSci*.

McClelland, J. L. 2009. The place of modeling in cognitive science. *Topics in Cognitive Science* 1(1): 11–38.

Milli, S.; Lieder, F.; and Griffiths, T. L. 2017. When does bounded-optimal metareasoning favor few cognitive systems? In *AAAI*, 4422–4428.

Noguchi, T.; and Stewart, N. 2014. In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition* 132(1): 44–56.

Noguchi, T.; and Stewart, N. 2018. Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological review* 125(4): 512.

Pettibone, J. C. 2012. Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making* 7(4): 513.

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. A.; and Botvinick, M. 2018. Machine Theory of Mind. In *International Conference on Machine Learning*, 4218–4227.

Rao, R. P. 2010. Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Frontiers in computational neuroscience* 4: 146.

Roe, R. M.; Busemeyer, J. R.; and Townsend, J. T. 2001. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review* 108(2): 370.

Ronayne, D.; and Brown, G. D. 2017. Multi-attribute decision by sampling: an account of the attraction, compromise and similarity effects. *Journal of Mathematical Psychology* 81: 11–27.

Russell, S. J.; and Subramanian, D. 1994. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research* 2: 575–609.

Savage, L. J. 1972. *The foundations of statistics*. Courier Corporation.

Simsek, O.; Algorta, S.; and Kothiyal, A. 2016. Why most decisions are easy in Tetris—And perhaps in other sequential decision problems, as well. In *International Conference on Machine Learning*, 1757–1765.

Spektor, M. S.; Gluth, S.; Fontanesi, L.; and Rieskamp, J. 2019. How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological review* 126(1): 52.

Stewart, N.; Chater, N.; and Brown, G. D. 2006. Decision by sampling. *Cognitive psychology* 53(1): 1–26.

Todd, P. M.; and Gigerenzer, G. E. 2012. *Ecological rationality: Intelligence in the world.* Oxford University Press.

Trueblood, J. S. 2012. Multialternative context effects obtained using an inference task. *Psychonomic bulletin & review* 19(5): 962–968.

Trueblood, J. S.; Brown, S. D.; and Heathcote, A. 2014. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review* 121(2): 179.

Tsetsos, K.; Moran, R.; Moreland, J.; Chater, N.; Usher, M.; and Summerfield, C. 2016. Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences* 113(11): 3102–3107.

Tversky, A.; and Simonson, I. 1993. Context-dependent preferences. *Management science* 39(10): 1179–1189.

Usher, M.; and McClelland, J. L. 2001. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review* 108(3): 550.

Vlaev, I.; Chater, N.; Stewart, N.; and Brown, G. D. 2011. Does the brain calculate value? *Trends in cognitive sciences* 15(11): 546–554.

Wang, J. X.; Kurth-Nelson, Z.; Kumaran, D.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Hassabis, D.; and Botvinick, M. 2018. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience* 21(6): 860.

Wang, Z.; Bapst, V.; Heess, N.; Mnih, V.; Munos, R.; Kavukcuoglu, K.; and de Freitas, N. 2016. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224* .

Wedell, D. H. 1991. Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17(4): 767.

Wollschlaeger, L. M.; and Diederich, A. 2020. Similarity, attraction, and compromise effects: Original findings, recent empirical observations, and computational cognitive process models. *The American Journal of Psychology* 133(1): 1–30.