RSMG2 Progress Report

Human Sequential Decision-Making Modelling: A Machine Learning Approach

Haiyang Chen

Supervisor: Prof. Andrew Howes and Prof. Jeremy Wyatt
Thesis Group Members: Prof. Ata Kaban and Prof. Jeremy Wyatt

June 18, 2018

Abstract

This report summaries the work of the last six months and proposes a plan for the next step. Deciphering decision making has a central role in the computational foundations of intelligence which converges the insight between computer science, cognitive science and neuroscience. We model human choice through three classes of reinforcement learning methods, value-based, policy-based and hybrid model, using only ordinal features of choice options. Then the performances of these models are compared on the different distribution of environment and with different ordinal observation capabilities. The results indicate that reinforcement learning models with ordinal observations generate behaviour that corresponds to human data and also make novel predictions about the order in which information is gathered. Ordinal observations pay a pivotal roll in the choice tasks and could help improve decision making. In the next step, visual attention and prospect theory will be studied to build a better human-like model and understand how humans choose, of which insights will benefit not only social science but also Artificial Intelligence (AI).

Contents

1 Introduction **3**
- 1.1 Decision-making modelling . . . 3
- 1.2 Reinforcement learning . . . 3

2 Reinforcement learning model **4**
- 2.1 Value-based method . . . 5
- 2.2 Policy-based method . . . 5
- 2.3 Actor-Critic method . . . 5
- 2.4 Performance of the models . . . 6
  - 2.4.1 The effect of the ordinal observation noise . . . 7
  - 2.4.2 The effect of the environment distribution . . . 10
- 2.5 Predictions of the models . . . 11

3 Discussion **13**

4 Future work **14**
- 4.1 Improving current model . . . 14
- 4.2 Attentional model . . . 14
- 4.3 Prospect theory . . . 14

5 Timetable **14**

1 Introduction

Deciphering decision making has a central role in the computational foundations of intelligence[1] which converges the paradigm between computer science, cognitive science and neuroscience. Sequential decision making involves taking a sequence of actions to achieve the goal, maximising the expected cumulative rewards[2]. While reinforcement learning (RL)[3] is such a general goal-driven learning algorithm which could learn to accomplish the goal efficiently without hand-crafted rules and explicitly programming. RL could be a promising potential approach to model human sequential decision-making problems.

1.1 Decision-making modelling

Humans frequently perform decision-making under uncertainty within a wide range in our daily life. Some choices are unthinking, such as what I will dress and eat in the morning. Others are of great importance from which university should I choose and what job should I take after graduation. The real-world decision-making problems should make trade-offs in precision and costs of computation[1] since it is impossible to avoid uncertainty with bounded cognitive and perceptual capabilities. Most standard theories of decision-making [4, 5, 6] assume that the decision-maker always chooses the optimal option which has the highest utility and the expected value of an option should be independent of the options presented in the past and now. Actually, human preferences are sensitive to context and sometimes changed by adding a third alternative. Contextual preference reversal has been used as an evidence[4, 7, 8] against the common view that human decision making is rational. Thus it has attracted a great deal of attention among the researchers studying human decision-making tasks, especially focusing on models of preference reversals[4, 9, 10, 11].
Decision Field Theory (DFT)[12], Leaky Competing Accumulator (LCA)[13] and their extended model[14] could explain preference reversals and demonstrate that people are not expected value maximizers. In recent works[15, 16], the state of art understanding of preference reversals is that human choices can be analysed based on computational rationality[17] and are a consequence of expected utility maximization given noisy observations. In the framework of computational rationality[17], maximizing expected utility is used to approximate the optimal choices[18] under bounded information-processing mechanisms.

1.2 Reinforcement learning

The latest model[15] is not a sequential model and there are still challenges to underlie the information processing mechanisms of human choices. Recent work[19] explores the feasibility of modelling sequential decision problems as a Partially Observable Markov Decision Process (POMDP) which provides a rigorous mathematical framework for decision-making modelling. RL[3] algorithm is a good application to POMDPs and provides an efficient model of human learning processing. Researches[20, 21] have shown that optimal behavioural strategies could be learned in decision-making tasks using RL which has made substantial progress in policy selection.
However, these contributions have not directly addressed the problem of preference inference and RL may easily lose its viability for large problems. The gener-

ality of POMDPs would lead to mass computation and large feature space for acquiring optimal strategies. Deep learning[22] approaches have made remarkable progress on the preference inference problem but do not directly address policy selection. Systems combining deep learning and RL, such as deep Q-network (DQN)[23] and asynchronous advantage actor-critic (A3C) [24], can successfully learn control policies in a range of different environments and achieve the higher level understanding of the interacted environment. Recent model [25] optimises strategies for visual search using DQN which is a solution to a POMDP. Currently, we model the rational choices by three classes of reinforcement learning method, Q-learning, Monte-Carlo-Policy-Gradient and Actor-Critic, to verify the feasibility of this approach. Future work attempts to model the attentional processing in a deep learning framework in which recurrent neural network(RNN) will be used to optimise the sequential decision-making processing over time. We try to understand how humans choose actions in the decision making task and find the way to help people make better decisions.

2 Reinforcement learning model

This section presents the work that has been done in the last six month, mainly including three classes of RL algorithms: value-based, policy-based and hybrid Actor-Critic method. Both of them are three main approaches to solving RL problems and are applied to model human preference choices using ordinal observation features. The performances of three kinds of models are compared on the different distribution of environment and with different ordinal observation noises in the preference choices tasks. Finally, the predictions about the information perceiving and processing are presented by using the model.
Each task has 3 options which are described in terms of two attributes, a random probability p and a random value v. According to the reported study [15], we set that the probabilities p are β distributed and the values v are t distributed. During each episode, the agent takes actions, which are 6 comparisons and 3 choosing, and gets rewards after took an action in each task. The probability of ordinal error is the probability that the relations are sampled uniformly random from the relation set. The reinforcement learning model has $4^9 = 4096$ states which is consisted of 6 elements representing the ordinal observation features $\{f(p_A, p_B), f(p_A, p_D), f(p_B, p_D), f(v_A, v_B), f(v_A, v_D), f(v_B, v_D)\}$. Each order relation has 4 kind of values, which indicate the relation is unknown, greater, equal and less which are defined as

$$f(v_A, v_B) = \begin{cases} none, & unknown \\ >, & v_A > v_B - \tau_v \\ \equiv, & |v_A - v_B| \leq \tau_v \\ <, & v_A < v_B + \tau_v \end{cases}$$

The cost to take an action is -1. The task is terminated when the agent makes a choice. The reward is 10 if the agent chooses the option with maximum expected value. Otherwise, it is -10. The value function or policy function is updated by rewards in each task according to the methods described in following section.

2.1 Value-based method

Value-based methods learn the value of actions and choose actions based on the estimating expected value by acting greedy. The Q-learning [26] is one of the most popular value-based RL algorithms and the most important breakthrough in the development of an off-policy Temporal-Difference (TD) learning algorithm[3]. One-step Q-learning is used in this report and defined by

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \cdot \underset{a}{max} Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Learning rate α is set as 0.1. The discount factor γ is set as 0.9. The action-value function Q directly approximates the optimal action-value function Q^* which is independent of the following policy. Thus it is known as off-policy control. The function Q is an empty table which is initialized to zeros before training. In each episode, the action is chosen from state-action pair using policy driven from the function Q by $\epsilon - greedy$ action selection. In order to enable early convergence, the ϵ descend from 0.9 to 0.05 uniformly in the first 500 episodes, thus the model could converge within 1000 episodes instead of 5000 episodes.

2.2 Policy-based method

Unlike value-based methods of which policy is based on a value function, policy-based methods directly learn a parameterized policy which could select actions. Policy gradient method changes the policy parameters in the way improving the performance, which is established

$$\nabla J(\theta) \propto \sum_s d^{\pi_\theta} \sum_a q_\pi(s, a) \nabla_\theta \pi(a \mid s, \theta)$$

where d^{π_θ} is the stationary distribution of Markov chain for on-policy π_θ. REINFORCE [27] algorithm, also known as Monte Carlo Policy Gradient, uses the complete return G_t from time t, which accumulate all the rewards until the end of the task, as an unbiased sample of value function $q_\pi(s, a)$. Whereas it is of high variance leading to slow learning because of Monte Carlo sampling. In order to reduce variance of policy gradient, advantage function $A^{\pi_\theta}(s)$ is used with an baseline function $B(s)$. The update rule of REINFORCE algorithm with baseline is

$$\theta \leftarrow \theta + \alpha^\theta \gamma^t A^{\pi_\theta}(S) \nabla_\theta ln \pi(A_t \mid S_t, \theta)$$

The state-value function $\hat{v}(S_t, \mathbf{w})$ could be a nature choice for the baseline. Thus the advantage function is

$$A_t{}^{\pi_\theta}(S_t) \leftarrow G_t - \hat{v}(S_t, \mathbf{w})$$

where the state-value function $\hat{v}(S_t, \mathbf{w})$ is updated by the rule

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \gamma^t A_t{}^{\pi_\theta}(S_t) \nabla_{\mathbf{w}} \hat{v}(S_t, \mathbf{w})$$

2.3 Actor-Critic method

A hybrid method has grown in popularity, known as actor-critic, which combine the benefits of policy based method with learned value function. Unlike the

Monte Carlo methods, such as REINFORCE , actor-critic tradeoffs variance reduction with bias introduction and thus accelerates learning. Policy function ('actor') learns the parameters from the feedback of the value function ('critic'). Instead of complete return of REINFORCE, one-step actor-critic algorithms update with one-step return as follows:

$$\theta \leftarrow \theta + \alpha^{\theta}\gamma^{t} A^{\pi_\theta}(S) \nabla_\theta ln\pi(A \mid S, \theta)$$

Where the state-value function $\hat{v}(S_t, \mathbf{w})$ use a learned state-value function as the baseline

$$A^{\pi_\theta}(S) \leftarrow R + \gamma\hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$$

The state-value function $\hat{v}(S_t, \mathbf{w})$ is updated by the same rule described above. The actor-critic method uses learned state-value function as a baseline and a critic. Thus it can significantly reduce the high variance of policy gradient and also introduce the bias of estimated value function.

2.4 Performance of the models

To predict the performance of the reinforcement learning models, we train the models in different environment distributions and with different ordinal observation noises. Then the models are used to predict the proportion of each choice and underlay the information perceiving processing.
In order to present the processing of training, we record three learning variables described as following:

- Training rewards: the sum of total rewards that the agent received after took sequential actions in 1000 training tasks. The reward of each option has a probability p of winning the value v in the form $p(v)$and $(1-p)0$. Shown in Figure 1 (top).
- Episode lengths: the total steps that the agent takes in 1000 training tasks or an episode. Shown in Figure 1 (middle).
- Episode Accuracy: the proportion of the chosen options, which have the highest expected value, using the latest value function or policy function in 100 testing tasks after each episode. The testing tasks are randomly sampled from the same distributions and different from the training data. Shown in Figure 1 (bottom)

The comparison of the results shows that REINFORCE and actor-critic method learn faster than the Q-learning method in the preference choice tasks, whereas they more likely converge to a local minimum. REINFORCE and actor-critic algorithms could converge after 200 training episodes while Q-learning algorithm tends to learn slowly leading to more than 600 training episodes. In other words, the policy based methods are data efficiency which could reduce the number of interactions with the environment. However the policy based methods, especially REINFORCE algorithm, may converge to a local minimum and have bad performance.

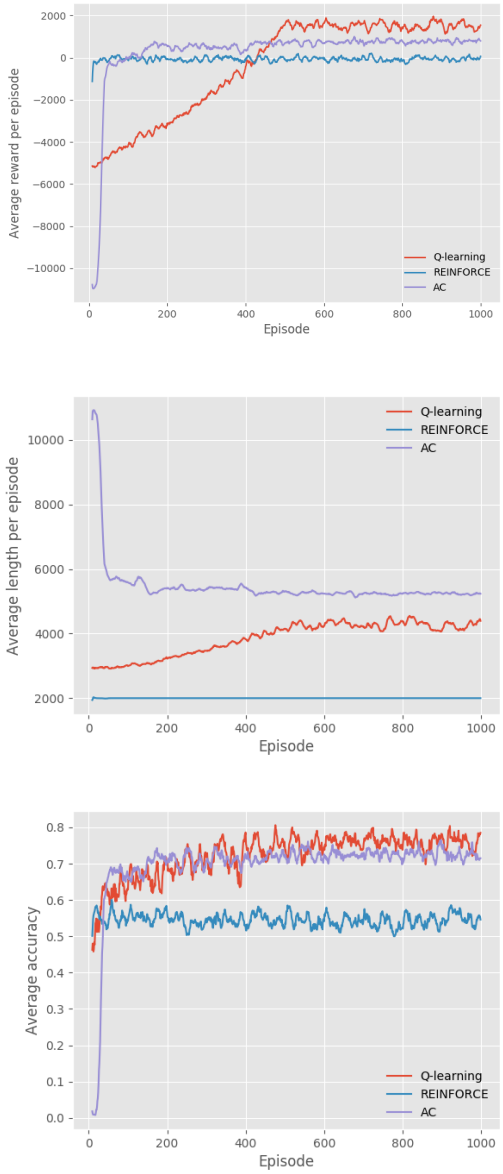

Figure 1: Comparison of performances of three RL algorithms in preference choice tasks. Learning curves for algorithms are training rewards, average lengths and accuracy over the episode from top to bottom. Every curve is smoothed with a moving average window of 10.

2.4.1 The effect of the ordinal observation noise

In each episode, the model is trained by 1000 tasks which are randomly sampled from β distribution $(a = 1, b = 1)$ for the probability p and $t-$distribution $(location = 19.60, scale = 8.08, df = 100)$ for the value v. There are 1000, 100,

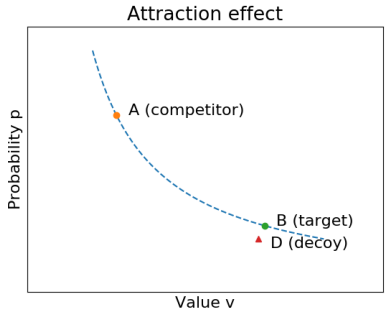

Figure 2: One type of contextual preference reversal–attraction effect in choice tasks. A has high probability while B has high value. A and B has equal expected value presented as the dotted line. A is the competitor and B is the target, which dominates decoy D on both attributes.

200 episodes in Q-learning, REINFORCE and actor-critic model. After a fixed number of training episodes, the model could converge and is tested by 200 choice tasks. For 100 testing tasks, the decoys are positioned closer to option A and the constraints are ($L1 = p_A > p_D > p_B, v_B > v_A > v_D$). For the other 100 tasks, the decoys are positioned closer to option B and the constraints are ($L2 = p_A > p_B > p_D, v_B > v_D > v_A$) shown as Figure 2. For each probability

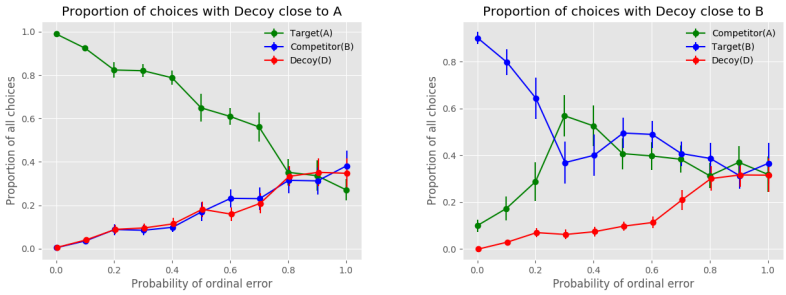

Figure 3: Predicted proportion of each choices by Q-learning agent.

of ordinal error, the simulation processing repeats 30 times for each Q-learning model and 10 times for each REINFORCE and actor-critic model which are more stable. The predicted proportion of choices against ordinal observation errors are shown in Figure 3, 4, 5. The left panel is when A is the target and the right is when B is the target.

The results show that RL models, both Q-learning and actor-critic, can perform preference reversals when the agents are able to choose the options with maximizing expected value. The preference reversals rate decreases as choices accuracy decreases and increases as choices accuracy increases. The ordinal observation noise causes the model to predict the reduced proportion of selections of the target and increased proportion of selections of the competitor, which is consistent with the human behaviour[28] and the reported model[15]. Comparison of Figure 3 and 5 indicates that actor-critic model is much more stable

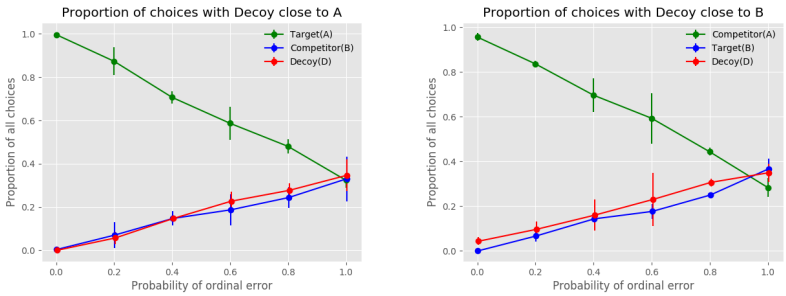

Figure 4: Predicted proportion of each choices by REINFORCE agent.

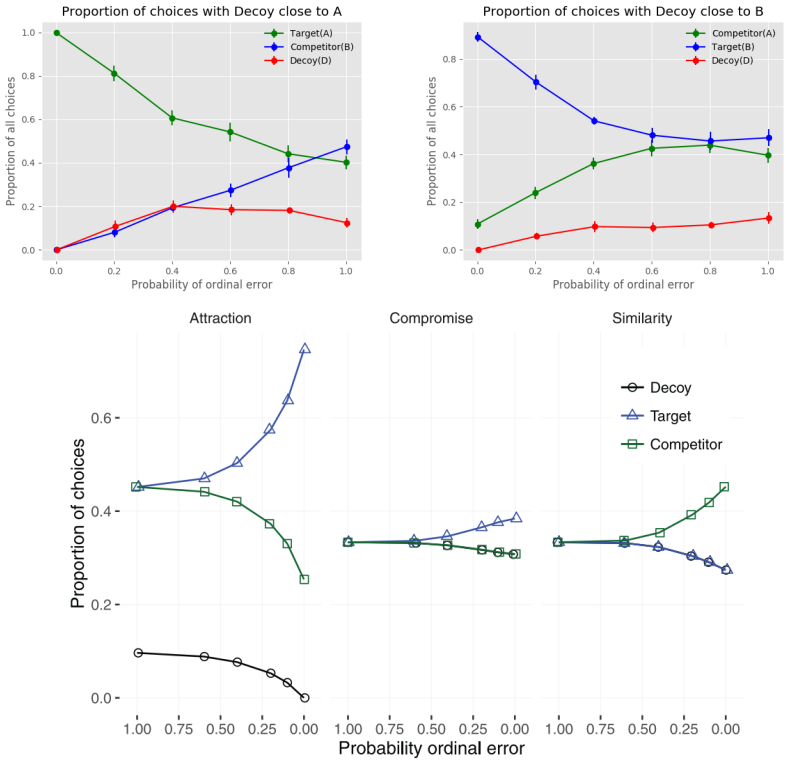

Figure 5: Top panel: Predicted proportion of each choice by AC agent. Bottom panel: Predicted effect of ordinal observation error on preference reversals (x-axis reversed), from 'Why contextual preference reversals maximize expected value'[15, Figure 11].

than Q-learning model. The REINFORCE agent cannot distinguish the situation when the decoy positions changes since it usually converge to local optimal early[3].

2.4.2 The effect of the environment distribution

The training is similar to the processing described above, but the environment distribution changed. For one kind of environment distribution, the simulation processing repeats the times same as above. The predicted proportion of choices against the location of the distribution of value v in the environment for the different level of ordinal observation noise (1 level in each row) and different level of scale (the lines in each figure) are shown in Figure 6.

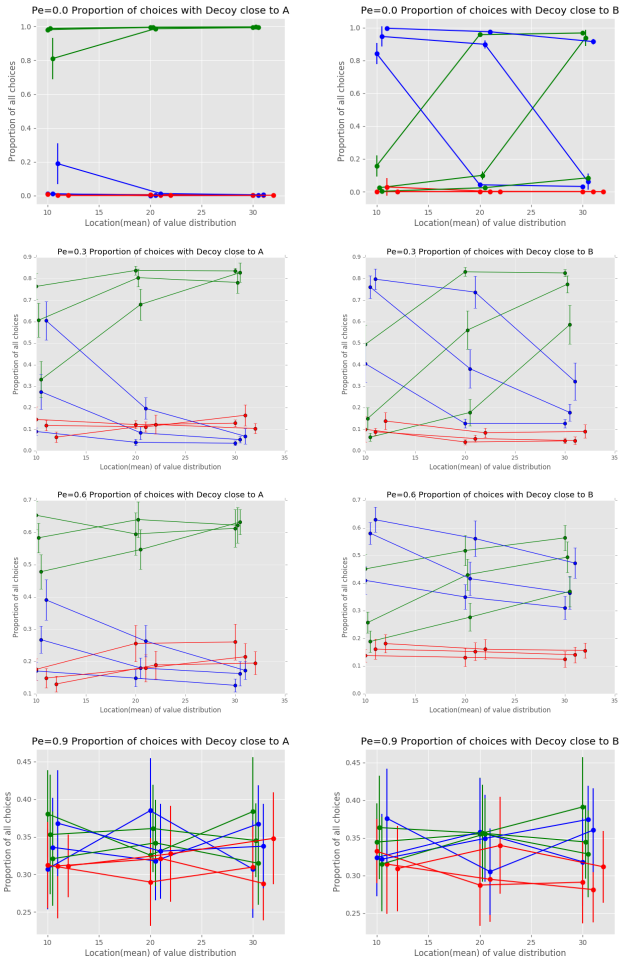

Figure 6: Predicted proportion of each choice by the Q-learning agent. The green colour represents the proportion of A, the blue colour represents B and the red colour represents D. In order to display the result clearly, the lines are plotted with x-axis shifting. From left to right, 3 lines are mapping 3 lever of scale $[scale = 4, 8, 12]$ for one location.

There are four levels of ordinal observed noise ($n = [0.0, 0.3, 0.6, 0.9]$) in each simulation. Figure 6 shows that the predicted effect for different levels of the scale and location of the environment distributions for the value v. The predicted proportion of choices varies in the testing tasks where the decoy posited.

The model could not distinguish the options with the maximum expected value from the others very well in some tasks, such as the decoys are positioned closer to option B when the parameters are set as $n = 0.0, scale = 4, location = 20$ and $n = 0.3, scale = 4/8/12, location = 30$. One possible reason is that the two attributes have different weights on the expected value of the options from different environment distributions. Overall, the Q-learning model could perform preference reversals in the different environment and with different ordinal observation errors, although the proportion changes by the different parameters. Thus we can fit the human data by tuning such the parameters.

2.5 Predictions of the models

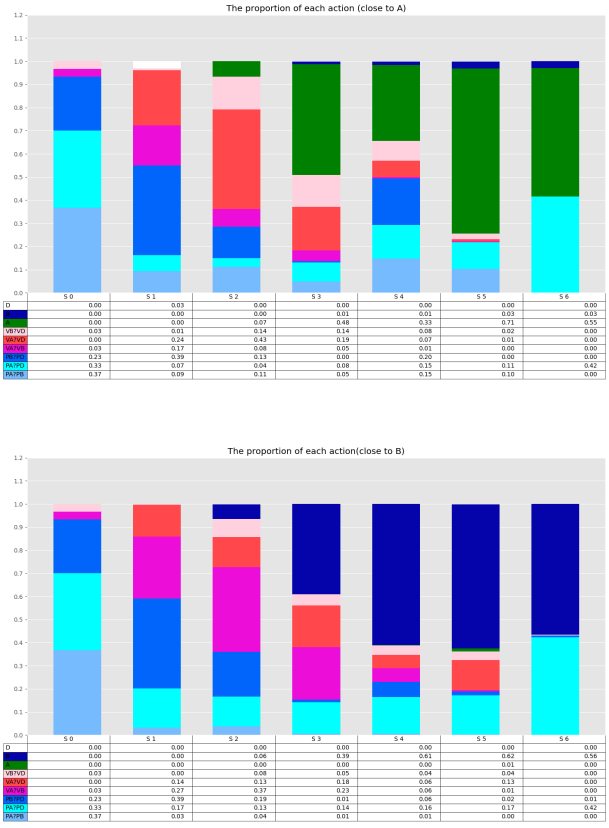

Figure 7: The predicted average proportion of each action taken in each step in choice tasks by the Q-learning agent. Comparison of the probability attributes between options is shown in blue colour: shy blue, cyan and bright blue. Comparison of the value attributes is shown in red colour: fuchsia, light red and light pink. Choosing A, B or C is shown in pure green, royal blue and pure red.

To explore the order in which ordinal information is gathered, the learned Q-learning models are used to predict the perceiving processing of ordinal features

when the parameters are set as $[n = 0.0, df = 100, scale = 4, location = 20]$. Figure 7 presents the average proportion of every ordinal relation used in each step during 100 tasking by 30 learned models which are trained in the same way described above.

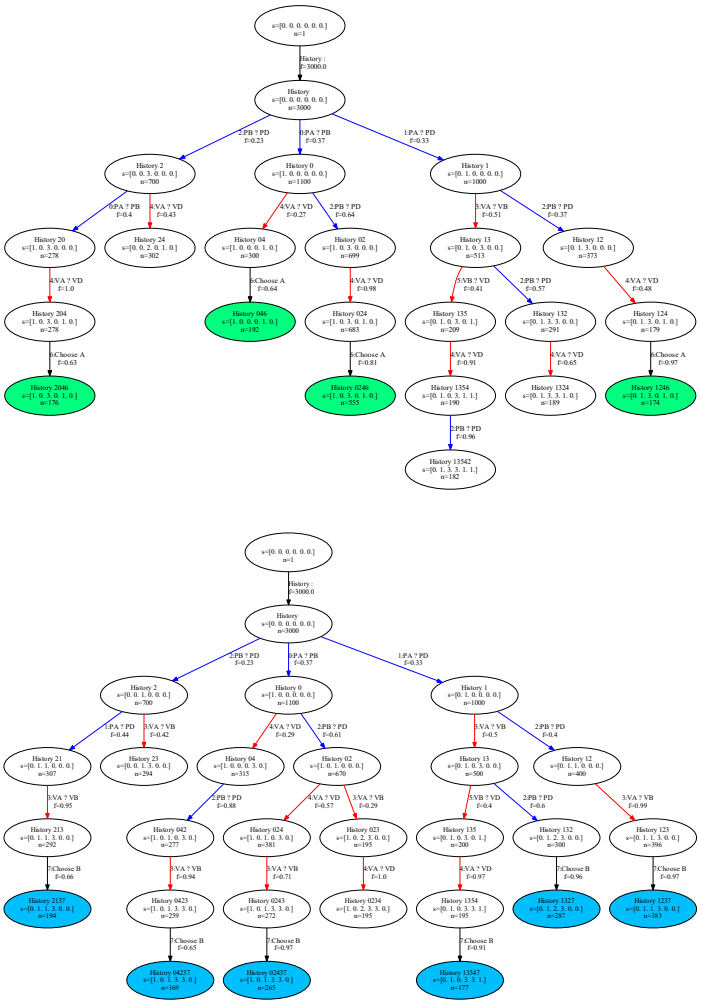

Figure 8: Predicted processing of choices by the Q-learning agent. Top tree: the situation when decoy posited close to A. Bottom tree: the situation when decoy posited close to B. To avoid the noise effect, the decision tree uses 95% confidence level. Comparisons of value attributes are presented in red arrows and comparisons of probability attributes are presented in blue arrows. Choosing A, B or C is shown in green, blue and red.

The results indicate that the agent mainly compares the probability attributes between options in the first few steps, then the value attributes and finally make a choice. The results are consistent with the observed human behaviour [29] that pair alternatives are compared on a single attribute dimension in each choice

using eye-movement data. The agent almost chooses targets in both situation although the perceiving processing order is the same. It is well supported that the model could adapt to different environments.
We also build two decision trees to record each state and actions occurred in totally 3000 tasks described above. Each node represents one state of the model and each arrow represents one action taken by the model. Each node contains 3 values which are action history from the start state, information of current state s and visit count n. Each action is coded by a number from 0 to 5. For example, ' History 1320456; s=[1, 1, 3, 3, 1, 1]; n=85' means that the agent has taken the following actions in order: compare the probability of A and D – '1: PA?PD', compare the value of A and B – '3: VA?VB', compare the probability of B and D – '2: PB?PD', compare the probability of A and B – '0: PA?PB', compare the value of A and D – '4: VA?VD', compare the value of B and D – '5: VB?VD', choose option A –'6: choose A'; the value of s means the order relation between the attributes; the visit count ' n=85' means this state occurs 85 times.
Figure 8 shows that the agent took $4 \sim 5$ comparisons before makes a choice and could choose the options having highest expected value mostly wherever the decoy D locates in. As we known, the amount of the most efficient perceiving actions are also $4 \sim 6$ which indicates that the model could learn an optimal way to gather the information without prior knowledge. The information perceiving processing shows the Markov property that the agent achieves the state 's=[1, 0, 3, 0, 1, 0]' and 's=[0, 1, 1, 3, 0, 0]' in different order whereas makes the same choice. It also shows that the agent uses the comparison of probability attributes, which are blue arrows, more than the comparison of value attributes, which are red arrows, in the first few actions. The comparison of value attributes is used more before making a choice.

3 Discussion

In summary, we have compared three classes of RL model in preference choice tasks. AC model is much more stable and of high date efficiency, however, it may converge on local minimum like the other policy gradient methods. The RL model with ordinal observations generates behaviour that corresponds to observed human data [29, 28] in different environments and situations without prior knowledge. The results indicate that the choices made by the model are rational when the model is able to choose the options with maximizing expected value mostly. It well supports the latest view[15, 16] on preference reversals that human decision making is optimal and a consequence of expected utility maximization given noisy observations. It also makes novel predictions about the ordinal features perceiving processing which can help to underlay the order in which information is gathered.
The state of art model[15] of context preference reversals demonstrates that human is expected value maximizer in preference choice tasks, however, it is not a sequential process model and analyses little about the information processing mechanisms. Our RL model is a general approach driven by the goal that maximizing the total expected cumulative values. It can predict the ordinal features perceiving processing in preference choice tasks. The neurocomputational process models of choice, such as DFT[12] and LCA[13] model, are a

sequential model that can make fast and reasonable choices and also present the processing of human choice. As discussion in[7], one limitation of the neurocomputational process model is that *"rational strategies of choice require the flexibility to modify the choice parameters in response to the environment and demands"*[7, p. 297]. In other words, it needs explicitly programming and designing the features to different situations and thus the neurocomputational work lacks flexibility and adaptivity. While the RL approach could learn approximating optimal policy in a general and flexible way without hand-crafted rules and explicitly programming like all the machine learning methods.
Currently, we investigate the feasibility of RL approach to model human choice. One potential limitation of the RL models reported here is that they are still built on selected futures and have not fitted observed human data well. In the next step, the combination of RL and deep learning, which can represent learning features and approximate function in complexity issues, could be a promising method to model human decision making in the more general way.

4 Future work

4.1 Improving current model

1. Improve the model to fit human data by adding calculating observation features, adjusting ordinal observation noise and environment distributions.

2. Implement asynchronous advantage actor-critic (A3C)[24] algorithm, which uses parallel multiple actor-learners, to improving training stability.

3. Try to use neural network approximating the value function or policy function—deep reinforcement learning (DRL) for generality and accuracy.

4. Try to build the experiment environment that using the panels, which is consisted of 6 pictures representing the information of 3 options described in the experiment[16], as the input of DRL model. Then explore whether the model could mimic human behaviour such as preference reversals.

4.2 Attentional model

Combine recurrent neural network(RNN) with reinforcement learning for visual search in complex images such as radiographs.

4.3 Prospect theory

Build the model under the foundation of prospect theory.

5 Timetable

End of June:

- Improving current model: add calculating observation features and tune parameters to fit human data. Implement A3C algorithm to improve training stability.

- Attentional model: study and implement simple RNN algorithm.

From July to August:

- Improving current model: build experiment environment and implement DRL algorithm.
- Attentional model: apply RNN method to model visual attention.

From September to October:

- Attentional model: use the model to solve problems in medical diagnosing.
- Prospect theory: study the prospect theory and build model based on it.

References

[1] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

[2] Ron Sun and C Lee Giles. Sequence learning: from recognition and prediction to sequential decision making. *IEEE Intelligent Systems*, 16(4):67–70, 2001.

[3] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[4] Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management science*, 39(10):1179–1189, 1993.

[5] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[6] Ivo Vlaev, Nick Chater, Neil Stewart, and Gordon DA Brown. Does the brain calculate value? *Trends in cognitive sciences*, 15(11):546–554, 2011.

[7] Marius Usher, Anat Elhalal, and James L McClelland. The neurodynamics of choice, value-based decisions, and preference reversal. *The probabilistic mind: Prospects for Bayesian cognitive science*, pages 277–300, 2008.

[8] Itamar Simonson. Mission accomplished: What's next for consumer bdt-jdm researchers? 2014.

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

[10] Lena Maria Wollschläger and Adele Diederich. The 2n-ary choice tree model for n-alternative preferential choice. *Frontiers in psychology*, 3:189, 2012.

[11] Jennifer S Trueblood, Scott D Brown, and Andrew Heathcote. The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, 121(2):179, 2014.

[12] Jerome R Busemeyer and James T Townsend. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432, 1993.

[13] Marius Usher and James L McClelland. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, 108(3):550, 2001.

[14] Robert M Roe, Jermone R Busemeyer, and James T Townsend. Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review*, 108(2):370, 2001.

[15] Andrew Howes, Paul A Warren, George Farmer, Wael El-Deredy, and Richard L Lewis. Why contextual preference reversals maximize expected value. *Psychological review*, 123(4):368, 2016.

[16] George D Farmer, Paul A Warren, Wael El-Deredy, and Andrew Howes. The effect of expected value on attraction effect preference reversals. *Journal of behavioral decision making*, 30(4):785–793, 2017.

[17] Richard L Lewis, Andrew Howes, and Satinder Singh. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2):279–311, 2014.

[18] Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

[19] Antti Oulasvirta, Per Ola Kristensson, Xiaojun Bi, and Andrew Howes. *Computational Interaction*. Oxford University Press, 2018.

[20] Daniel Acuna and Paul R Schrater. Structure learning in human sequential decision-making. In *Advances in neural information processing systems*, pages 1–8, 2009.

[21] Xiuli Chen, Gilles Bailly, Duncan P Brumby, Antti Oulasvirta, and Andrew Howes. The emergence of interactive behavior: A model of rational menu search. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4217–4226. ACM, 2015.

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[23] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[24] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

[25] Aditya Acharya, Xiuli Chen, Christopher W Myers, Richard L Lewis, and Andrew Howes. Human visual search as a deep reinforcement learning solution to a pomdp. 2017.

[26] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

[27] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer, 1992.

[28] Jennifer S Trueblood, Scott D Brown, Andrew Heathcote, and Jerome R Busemeyer. Not just for consumers: Context effects are fundamental to decision making. *Psychological science*, 24(6):901–908, 2013.

[29] Takao Noguchi and Neil Stewart. In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1):44–56, 2014.