Modelling Emotion Based Reward Valuation with Computational Reinforcement Learning

Can Koluman (can.koluman@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Christopher Child (C.Child@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Tillmann Weyde (T.E.Weyde@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Abstract

We show that computational reinforcement learning can model human decision making in the Iowa Gambling Task (IGT). The IGT is a card game, which tests decision making under uncertainty. In our experiments, we found that modulating learning rate decay in Q-learning, enables the approximation of both the behaviour of normal subjects and those who are emotionally impaired by ventromedial prefrontal lesions. Outcomes observed in impaired subjects are modeled by high learning rate decay, while low learning rate decay replicates healthy subjects under otherwise identical conditions. The ventromedial prefrontal cortex has been associated with emotion based reward valuation, and, the value function in reinforcement learning provides an analogous assessment mechanism. Thus reinforcement learning can provide a good model for the role of emotional reward as a modulator of the learning rate.

Keywords: reinforcement learning; Q-learning; learning rate decay; Iowa Gambling Task; ventromedial prefrontal impairment

Introduction

According to psycho-evolutionary theorists, emotions assist the organism in maintaining homeostasis relative to its behavioural and survival goals (Plutchik, 2003). The emotion feedback mechanism solves problems without the need for higher cognitive analysis (Damasio, 2006).[1] Rolls (2013, Ch. 4) proposes that emotions regulate instrumental learning and influence contingent outcome-action selection.

The pre-frontal cortex and its regions play a key role in goal directed learning and behaviour (Miller & Cohen, 2001). Ventromedial prefrontal cortex (VMF) lesions produce a characteristic learning deficit, where the subject, while retaining good intellectual function and understanding, is no longer able to learn from real life mistakes. Wallis (2007) has argued that the VMF provides emotion valuation input critical for good decision making.

The Iowa Gambling Task (IGT) was the first clinical test, which identified VMF impairment in human trials (Bechara, Damasio, Damasio, & Anderson, 1994). In the IGT, subjects need to choose a card from one of four decks. There are two 'good,' and two 'bad' decks, but the 'bad' decks start with positive rewards. Once penalties set in on the bad decks, subjects should adjust the choice of decks accordingly. Fellows and Farah (2005, 2004) present a re-shuffled variation of the original IGT, where penalties start earlier in the bad decks. While VMF impaired subjects fail the original IGT, they pass the re-shuffled variant. On the basis of these differing test results, Fellows and Farah (2005, 2004) link VMF impairment to reversal learning deficit.

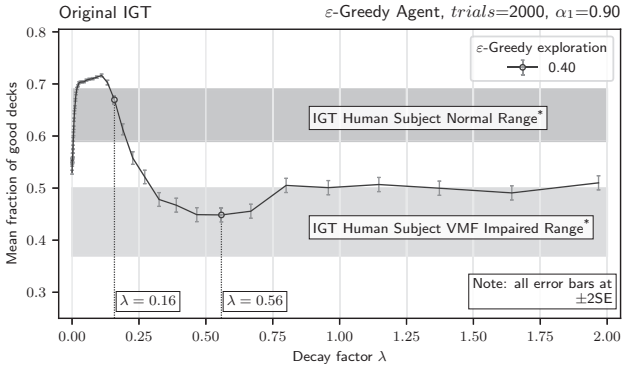

Figure 1: The ε-Greedy agent applied to the original IGT data with different learning rate decay values. A low decay rate yields normal behaviour, whereas a high decay rate reproduces VMF impairment. *See text and Table 4 for details.

Computational reinforcement learning methods approximate an optimal decision policy by iteratively aggregating time-contingent reward values (Sutton & Barto, 2018). For example, reinforcement learning techniques may be used to calculate a suitable path for escaping a maze (Osmankovic & Konjicija, 2011).

Watkins (1989) developed, the Q-learning framework, a reinforcement learning model, which, in addition to the discount rate, uses a single novel parameter denoted by α, known as the learning rate. The learning rate determines the relative contribution of current yield to accumulated value. We add to the Q-learning model a decay factor parameter λ, which produces exponential decay of the learning rate (Powell, 2011, pp. 427). We show that Q-learning, with the addition of learning rate decay, reproduces the clinical results of the original and re-shuffled IGT variants.

[1] First published in 1994 by G.P. Putnam's Sons, New York, USA.

Figure 1 presents our key result for the original IGT variant. For low learning rate decay, the ε-Greedy agent performs in the range of normal human subjects. As learning rate decay increases, agent performance reduces to the range of VMF impaired subjects. The dark and light gray zones mark the mean fraction of cards chosen from the good decks, reported in the literature for normal and VMF impaired subjects respectively.

In the remainder of this paper, we first review related literature. Then we discuss the IGT in detail, develop the computational treatment, and summarize the experimental design and the results. These are followed by the discussion and conclusion.

Literature Review

According to Dalgleish (2004), the prefrontal cortex constitutes a primary anatomical locus for animal and human behaviour attributed to emotion. Dalgleish's prefrontal cortex includes the areas (Krawczyk, 2002, pp. 633-635) others have called the VMF (Bechara, 2004) or OFC (Rolls, 2000; Wallis, 2007). This paper uses the term VMF to refer to the area of the prefrontal cortex involved in valuation by emotion. However, some studies prefer the term orbitofrontal cortex, or OFC. This section retains the respective authors' original use of the terms OFC or VMF.

VMF impaired patients, can recognise poor decisions and describe good decision making strategies, but exhibit a distinctive inability to learn from their mistakes (Bechara et al., 1994). In IGT studies, this inability applies to negative (Bechara et al., 1994) and positive rewards (Bechara, Tranel, & Damasio, 2000).

To explain VMF impaired deficits, Damasio (1998) proposes the Somatic Marker Hypothesis: an involuntary feedback mechanism where a physical or virtual body sensation is associated with a particular emotion. VMF impairment disrupts somatic marker pathways, and the affected individual remains in a slow, logic based decision making paradigm (Bechara, 2004; Damasio, 1998, 2006). Others have instead advanced the view that VMF impairment leads to loss of reversal learning ability (Dunn, Dalgleish, & Lawrence, 2006; Maia & McClelland, 2005; Fellows & Farah, 2003, 2005, 2004). Reversal learning ability is the facility to unlearn a stimulus-response-association, which had previously produced favourable emotion-valued outcomes.

The VMF is also associated with emotion (Krawczyk, 2002; Hornak et al., 2003; Rolls, 2000). Modelling emotion in learning and decision making has been challenging (Volz & Hertwig, 2016). Without using emotion, the Rescorla-Wagner classical conditioning model presents a learning rule for assessing the pre and post trial associative strength of a new stimulus (Rescorla & Wagner, 1972). TD(λ) reinforcement learning methods extend the Rescorla-Wagner model and enable intra-trial assessment of an associative stimulus (Sutton & Barto, 2018, pp. 350-357). Contingent stimulus-response animal studies also inspired Q-learning. However, unlike Rescorla-Wagner, Q-learning does not explain the conditioning mechanism, but instead develops a decision theoretic learning framework (Watkins, 1989). Q-learning remains one of the most successful machine learning algorithms, especially as the feedback stage for deep neural networks (Mnih et al., 2015).

Puviani and Rama (2016) propose a complex, neurologically motivated emotion learning framework, which models both the OFC and the Amygdala. However, typically computational emotion synthesis employs more abstract, behaviourally driven approaches based on varied psychological views. Recently, reinforcement learning approaches incorporating emotion have been receiving increased attention. Reinforcement learning can produce lightweight models, has close ties to optimal control, and provides an intuitive approach for aggregating contingent values (Powell, 2011; Sutton & Barto, 2018).

Moerland, Broekens, and Jonker (2018; 2017) identify and survey 52 papers published from 1998 to 2016 relating to emotion and reinforcement learning. They report four common methods for eliciting emotion: homeostatic targets, introspective appraisal, value function or reward modulation, and, sensor or sense driven. Emotions influence rewards, contingencies, modulate the exploitation versus exploration trade-off, and sometimes directly act on action selection. Typically, the value function itself aggregates emotion modulated inputs into an action selection mapping. We believe that emotion modulated reinforcement learning thus aims to encapsulate the functionality of the VMF.

While developing our model, the Moerland et al. (2018; 2017) survey had not yet come out. However, we had considered Broekens, Jacobs, and Jonker (2015), where joy, distress, hope, and fear act as value inputs into TD(0) computational reinforcement learning. In contrast to Moerland et al. (2018; 2017) and Broekens et al. (2015), our model does not need an emotion generation layer. In the context of the discussed models, our model re-interprets the Q-value function as a single aggregated emotion signal. While our learning rate is modulated by another hyper-parameter, the decay factor, we do not synthesize emotions to modulate these hyper-parameters. Instead, we use an external search grid to assess the end-effect of learning rate changes, which we hypothesize might result from VMF impairment.

Our learning rate decay law does not satisfy the well-known statistical convergence requirement that the sum of the learning weights must be infinite (Robbins & Monro, 1951; Spall, 2003). In practice, fully proving theoretical statistical convergence is difficult (Spall, 2003, p. 122), and proof of theoretical convergence does not automatically ascertain good model performance (Powell, 2011, p. 450). Moreover, an individual organism and its decision making mechanisms possess a finite lifespan. Therefore we think it is valid to investigate finite term, periodic decisions with tools where statistical convergence is not theoretically guaranteed. We propose that our method of simulating human behaviour with

learning rate decay could form a useful baseline for generalised reinforcement learning solutions. We focus here on the empirical effect of learning rate decay on decision quality and learning.

The Iowa Gambling Task

The original (Bechara et al., 1994) and re-shuffled (Fellows & Farah, 2005, 2004) Iowa Gambling Task (IGT) variants form the basis of this paper and we explain them here in more detail.

Description

The IGT is a card game where the participant receives a loan, and should maximize profit including repayment of any loans. The card game consists of four decks: A, B, C, and D. The participants are told that "some decks are worse than others." (Bechara et al., 2000, p. 2192) In each turn, the participant draws one card from any deck. For each draw, the participant then receives a fixed reward, and occasionally has to pay a fine. Decks C and D, known as the 'good' decks, give low fixed rewards, low fines, and, on average, yield net gains. The remaining two 'bad' decks, A and B, produce high rewards, but even higher losses, and, on average, produce a net loss.

The game stops after 100 turns, when the dealer announces the end. However, the participant does not know when the game will end. If the participant runs out of money, additional loans are available. The hypothesis is that the participants discover the 'good,' low risk decks and choose accordingly. A score of more than 50 draws from the good decks is defined as a normative pass by Fellows and Farah (2005, 2004).

While the original IGT lasts 100 turns, Bechara et al. (1994) only predefine a 40-draw sequence for each deck. They do not discuss whether any participants drew more than 40 cards from the same deck, and in the provided example draws, human participants do not draw more than 40 cards from the same deck. In our implementation, we use the published 40-draw predefined sequences. However, to ensure that a software agent could potentially draw more than 40 consecutive cards from the same deck, we loop at the end of each deck to the beginning of the deck.

Original and Re-shuffled Card Deck Differences

In the original IGT, the 'bad' decks, A and B, each start with an eight card long 'special' sequence, where the player receives positive net gains. Consequently, at the beginning of the task, the 'bad' decks appear 'good.' However, in each bad deck, the 'special sequence' is immediately followed by one or more high fines, causing the player, on subsequent selections, to lose all gains and move into debt.

In the re-shuffled variant, Fellows and Farah (2005, 2004) move the first 8 cards in each original deck to the end. This removes the initial confounding conditioning sequence, and players experience, across all decks, fines relatively quickly.

The full details of the original and re-shuffled decks can be found in Bechara et al. (1994, p. 9) and Fellows and Farah (2005, 2004, p. 59) respectively.

ε-Greedy Q-Learning with Learning Rate Decay

This section motivates and develops our Q-learning model with learning rate decay.

Computational Background

The IGT constitutes a version of the n-armed bandit problem (Ross, 1983, pp. 131-151): there are four processes, of which only one can be operated at any one time. The software agent devises a policy for gaining information (exploring), for assessing (scoring), and then choosing the most advantageous process (exploiting). Kuleshov and Precup (2000) present various classic computational techniques for scoring, and for balancing exploration versus exploitation. We employ Q-learning because it is simple and permits investigation of learning rates which vary from $1 \bullet n$ and its derivatives.

Single State Q-learning

We model the IGT as a single state environment with four card decks and four actions. We do not fully implement Q-learning as proposed by Watkins (1989) where the current contribution to the Q-factors uses off-policy updating. Instead, we apply on-policy value function updates as suggested by Sutton and Barto (2018, p. 32).

Given an action a, let $Q \bullet a \bullet$ be an unknown value function, and let $Q_n \bullet a \bullet$ denote the n^{th} iterative approximation. Then we write the computational estimation problem as:

$$Q_n \bullet a \bullet \bullet \ \alpha_n r_n^a \bullet \ \bullet 1 - \alpha_n \bullet \gamma Q_{n-1} \bullet a \bullet \tag{1}$$

where $r_n^a \bullet \ reward_n^a - fine_n^a$ is the net reward for action a at iteration n , γ is the discount rate, and α_n is the learning rate at iteration n. The discount rate γ, when set to less than 1, is used to devalue future yields r_n^a. We assume that the length of the card game, although unknown, is not long enough to create a preference for present rewards. Consequently, we set $\gamma \bullet \ 1$.

Learning Rate Decay

A rapidly decaying learning rate sequence, $\{\alpha_n\}$, can get close to 0 prior to some final period T and effectively curtail learning. We consider a geometric-decay learning rate sequence of the form (Powell, 2011, pp. 427):

$$\alpha_n \bullet \ \Lambda \alpha_{n-1} \tag{2a}$$

$$\Lambda \bullet \ 2^{-\lambda \bullet ln2} \tag{2b}$$

where $\lambda \in \left[0 \bullet \infty\right)$ is the *decay factor*, and $ln2$ is a normalizing constant used to rescale to natural logarithms in the computations.

Given equation (2b), $\{\alpha\}_n$ only satisfies the theoretical statistical convergence requirement $\sum_n \alpha_n \bullet \ \infty$ (Powell, 2011, pp. 274-285), when $\lambda \bullet \ 0$.

However, equations (2a) and (2b) always guarantee, in a finite number of iterations, computational convergence in the sense of $|Q_n - Q_{n-1}| \bullet \ \varepsilon$ for some $n \ll \infty$ and $\varepsilon \bullet \ 0$. In practice, our approach produces good approximations to normal as well as VMF impaired behaviour.

Table 1: Methodology, Simulation Parameter Summary

Trials, N	2000
Initial Learning Rate, α_1	0.05 to 1 by 0.05 steps
*Decay Factor, λ	$\lambda_i \bullet \lambda_{max} 2^{-ir \bullet ln2}$
ε-Greediness, ε	0.00 to 0.50 by 0.10 steps

*With $\lambda_{max} \bullet$ 3•3765, $r \bullet$ 0•012, $i \bullet$ 0•1•2••••

Table 2: Original IGT Test, Pixel Match Computed Means $\pm$ SE for Fraction of Cards Chosen from the Good Decks reported in the IGT Literature

Subjects	Study	N	Mean fraction of good decks
Controls	Bechara et al. (1994)	44	0.69 ± 0.015
	Bechara et al. (1998)	21	0.62 ± 0.032
	Bechara et al. (2000)*	20	0.59 ± 0.019
	Farah et al. (2004)	14	0.63 ± 0.023
VMF Impaired	Bechara et al. (1994)	6	0.37 ± 0.055
	Bechara et al. (1998)	9	0.40 ± 0.035
	Bechara et al. (2000)*	10	0.45 ± 0.028
	Farah et al. (2004)	9	0.50 ± 0.020

*Results reported in 20 draw blocks. Calculation of 100 draw values assume no inter-block covariance.

The ε-Greedy Agent

For most of the time, the ε-Greedy agent exhibits unconstrained maximizing behaviour, and at any iteration n, picks the deck with the highest attributed value:

$$Q_n^* \bullet \max_a Q_n \bullet a \bullet\bullet a \in \{A \bullet B \bullet C \bullet D\} \quad (3)$$

To ensure exploration, occasionally the ε-Greedy agent chooses an action randomly. Consequently, the agent's decision making rule is:

$$Q_{n \bullet \varepsilon}^* \bullet \begin{cases} Q_n^* \bullet & \text{with probability } 1 - \varepsilon \bullet \\ \text{choose } a \text{ randomly} & \text{with probability } \varepsilon \end{cases} \quad (4)$$

where $\varepsilon \in \bullet 0 \bullet 1 \bullet$ indicates the probability of exploration.

Experimental Design and Results

Simulations consist of multiple trials of 100 draws. All cross-section comparisons are conducted at the 100^{th} draw, which corresponds to the duration of the clinical tasks. Table 1 summarizes the parameter values used in this paper. We assess the parameter space with brute-force, grid-based searches.

As the original test data (Bechara et al., 1994, 2000; Fellows & Farah, 2005, 2004; Bechara, Damasio, Tranel, & Anderson, 1998) was not available, we converted the graphical presentations into numerical format using pixel matching. For each study, Tables 2 and 3 summarize, for normal and VMF impaired subjects, the pixel match calculated original and re-shuffled IGT test results respectively, reported in terms of the fraction of cards chosen from the good decks.

Table 3: Re-shuffled IGT Test, Pixel Match Computed Means $\pm$ SE for Fraction of Cards Chosen from the Good Decks reported in the IGT Literature

Subjects	Study	N	Mean fraction of good decks
Controls	Farah et al. (2004)	17	0.72 ± 0.038
VMF Impaired	Farah et al. (2004)	9	0.67 ± 0.078

Table 4: Original and Re-shuffled IGT Mean Fraction Good Deck Ranges Used for Comparing ε-Greedy Agent and Literature Results

IGT Variant	Original	Re-shuffled
Pixel matched studies	4	1
Comparison Rule	Table 2 Minimum and Maximum	Table 3 $\pm$ 2 SEs
Normal Match Range	0.59 to 0.69	0.64 to 0.80
VMF Impaired Match Range	0.37 to 0.50	0.51 to 0.83

Table 4 shows the pixel matched ranges of fraction of good decks we derived from IGT literature results and use to compare to the ε-Greedy agent results.

Results

We found that, given appropriate standard values for initial learning rate and exploration, learning rate decay λ proves to be the key variable, which determines the ε-Greedy agent's degree of success. We first present the results obtained from learning rate decay and exploration variations, and then discuss the effects of the initial learning rate.

The Effects of Learning Rate Decay and Exploration Fig. 2 shows, given exploration, the strong effect of learning rate decay on mean fraction of good decks. For the original IGT, as the decay factor increases, the mean fraction good decks achieved by the agent decreases; and, eventually approaches a value close to or below 0.5, the IGT fail criterion. But for the re-shuffled IGT, as the decay factor increases, mean fraction of good decks scores remain above 0.5.

Figure 2 also shows that for the original and re-shuffled decks, at $\varepsilon \bullet$ 0•40, the ε-Greedy agent matches actual IGT test subject behaviours: control subject behaviour is matched at a learning rate decay factor of $\lambda \bullet$ 0•16 (15% per period learning rate decay), and VMF impaired subject behaviour is matched at $\lambda \bullet$ 0•56 (43% decay).

$\varepsilon \bullet$ 0•40 constitutes the first exploration value at which we obtain a match for healthy and VMF impaired human performance zones. Further match candidates exist for $\varepsilon \bullet$

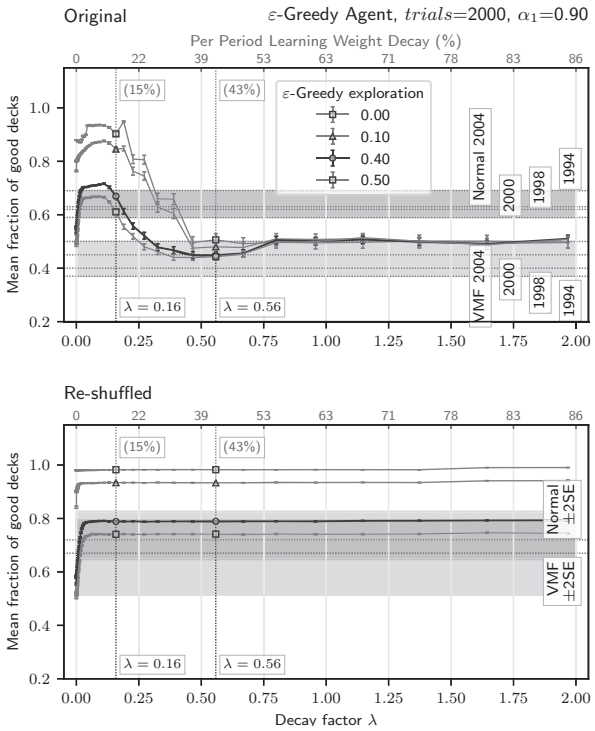

Figure 2: Learning rate decay and ε-Greedy agent IGT performance. Dark lines approximate human IGT behaviour. With high exploration, for the Original IGT, at the lower decay factor, the agent matches control subject performance; and, at the higher decay factor, the agent matches VMF impaired subject results. With high exploration, for the re-shuffled IGT, at the lower and higher decay factors, the agent matches human subject performance.

0•40 – 0•50, where the values of the agent's mean fraction of good decks are inside the match ranges for the corresponding values reported in the literature for human subjects.

At ε • 0•50, the agent explores 50% of the time. 50% exploration seems high. However, it constitutes a targeted strategy, for example, compared to always choosing lottery numbers randomly. We can also see that agents, which do not explore at all (ε • 0•0), or explore just a little (ε • 0•10), substantially exceed human performance. We discuss this result later.

Table 3 shows that the re-shuffled deck VMF impaired match range is derived from a single study with 9 participants. In Fellows and Farah (2005, 2004, p. 60, Figure 4), VMF impaired subject performance includes a high performace cluster of 3 subjects with a pixel matched cluster mean of 0.95. These 3 VMF impaired subjects achieve a re-shuffled deck test result approximated by the performance of our ε=0.10 agent, which achieves across all decay factors a mean fraction of good decks score of 0.92.

Having only a single re-shuffled deck study makes interpreting the statistical context of this high performance cluster difficult. Therefore in Table 4, we construct re-shuffled deck VMF performance match ranges using ± 2 standard errors, which produces approximately a 92% confidence interval (two-sided p-value: 0.080516). Our match range can be interpreted as the smallest match range based on the availability of a single study.

With re-shuffled decks, the decay factor λ influences the mean fraction of good decks by very little. This result appears to be driven by card sequencing. To test the effect of card sequencing, we created a new deck environment, where cards are drawn randomly, without replacement, from the original IGT decks. This new random draw card environment produces plots, which display a pattern similar to that of the original decks in Figure 2, except that as the decay factor increases, mean fraction of good decks decreases towards but remains above 0.5. Therefore relative to randomly ordered decks, both the original and re-shuffled decks create sequencing biases, which put different demands on learning: the original decks tax re-learning, while the re-shuffled decks teach via 'early punishment.' It would be interesting to test whether both normal and VMF impaired subjects pass the random draw version of the IGT as predicted by our simulation.

Finally, increasing exploration leads to a steady downward shift of the mean fraction of good decks plots with little effect on contour shaping. In contrast, learning rate decay λ appears key for determining agent behaviour; and increasing learning rate decay approximates the behaviour of normal and VMF impaired IGT participants.

The Effects of the Initial Learning Rate Unlike learning rate decay, the initial learning rate α_1, like exploration, only has a mild effect on the mean fraction of good decks.

Figure 3 shows the effect of the initial learning rate α_1 on mean fraction good decks at the 100th draw for the ε-Greedy agent with ε • 0•40. For the the original and re-shuffled decks, mean fraction of good decks scores vary little along the initial learning rate axis. In contrast, increasing learning rate decay leads to normative IGT fail (i.e., mean fraction of good decks ≤ 0.50) for the original decks; but not for the re-shuffled decks, thereby inducing agent behaviour to match human trial performance.

Discussion

In our Q-learning IGT simulations, learning rate decay λ constitutes a critical parameter. Increasing learning rate decay generates the observed behaviour of human IGT participants. For low learning decay factors, the ε-Greedy agent passes both the original and re-shuffled IGT. As we increase the learning decay factor, the agent fails the original test, while continuing to pass the re-shuffled variant. Therefore, increasing the decay factor leads to the learning behaviour of VMF impaired IGT participants.

In reinforcement learning, the software agent's internal valuation produces action selection. Rolls and others have ar-

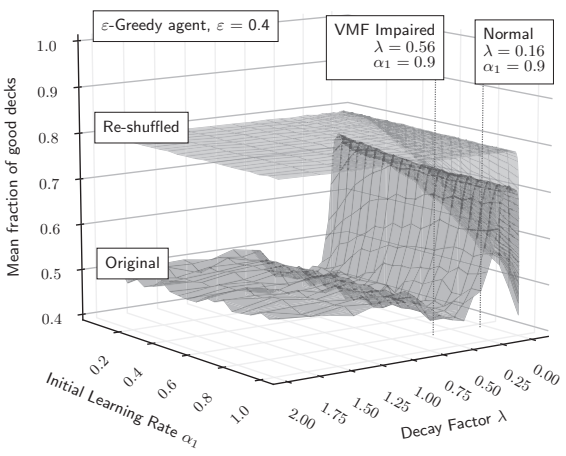

Figure 3: ε-Greedy agent with ε • 0•40: fraction of good decks by initial learning rate α_1, and decay factor λ. Initial learning rate α_1 variations exert a mild influence, while learning rate decay λ variations exert a strong influence on mean fraction of good decks. Normal and VMF impaired match values are marked accordingly for ε • 0•40.

gued that emotions result from reward assessment in the VMF (Krawczyk, 2002; Hornak et al., 2003; Rolls, 2000, 2013). We draw parallels between VMF provided reward values and the reinforcement learning process. The Q-value function encapsulates reward information. Learning rate decay λ can elicit progressive decay in current reward contribution. If learning rate decay is very high, then current reward value contribution decreases rapidly, and this leads to quick computational convergence. This effect produces two impediments, which may mimic VMF impairment: the value function not only 'finalises' too quickly, but also is itself dominated disproportionately by initial experiences.

Consequently with high learning rate decay, early and high 'bad' deck payoffs in the original IGT produce an incorrectly learned policy response: the 'bad' decks appear to be good. The ε-Greedy agent's beliefs, once established, even when presented with current information to the contrary, can no longer be modified. If emotion impairment due to VMF lesions removes the ability to unlearn previously learned responses, then in reinforcement learning, this behavioural effect can be achieved via high learning rate decay.

Conclusion

Bechara et al. (1994, in title) state that VMF impaired patients suffer from an "insensitivity to future consequences." Our simulated VMF impaired original IGT results suggest that this insensitivity comes from remaining mired in the past, and appears consistent with loss of the ability to reverse learning.

Interestingly, at lower exploration values, the ε-Greedy agent achieves mean fractions of good decks that are better than those achieved by human subjects. To match actual test subject behaviours, exploration has to be set at a high level.

It is not clear why agent behaviour, while qualitatively mirroring human behaviour, achieves better than human results. A number of possibilities could explain this finding. A reformulated model with decaying ε-Greediness may provide additional insight into the exploration versus exploitation tradeoff. Human behaviour may initially have higher exploration, which then progressively decreases with learning. In this paper, to keep the parameter count low, to avoid over-fitting, and to focus on the decay factor λ, we have not added any additional parameters for modelling variable exploration.

Alternatively, given the lack of full-knowledge, human behaviour may be more cautious. Human level learning has evolved for a wide variety of tasks, and therefore may perform optimally at other tasks for which Q-learning is less well suited. In contrast, grid search allows the searcher to become all-knowing with respect to the parameter space. For humans with incomplete information, keeping exploration high may make sense, just in case a deck would produce some unexpected yields later in the task.

Finally, it is also possible that the calculations performed by reinforcement learning agents are too hard for mental arithmetic and that the lack of precise calculations leads to sub-optimal decisions.

In a psycho-evolutionary context, emotions provide a flexible mechanism for establishing homoeostatis under environmental uncertainty (Plutchik, 2003; Rolls, 2013). If this environmental uncertainty fulfils certain regularity conditions, such as distributional full, or bounded, time-invariance, existence of the mean, or high-yield state correlation, then there could be high survival value to speculative learning; that is, deriving a working decision policy from just a few samples. From short learning bursts, the organism, or agent, could converge, to a long-term optimal decision rule. Emotions (via learning rate decay) could be responsible for opening and closing a short learning window. It is possible that the VMF driven emotion mechanism has evolved to produce the ability for organisms to learn efficiently from just a few samples.

Humans have evolved as generalised decision learners. In many machine learning tasks, only a narrow range of hyperparameter values produce a coherent result. Therefore the addition of a learning decay factor, which mimics human learning could provide an ideal starting point over a number of tasks for computational learning. Overall, our results indicate that computational reinforcement learning may be used as the basis for modelling emotion based learning. The results are encouraging for further investigation into more complex forms of learning and emotions.

References

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition*, *55*(1), 30–40.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15.

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation Of Working Memory from Decision Making within the Human Prefrontal Cortex. *Journal of Neuroscience*, *18*(1), 428–437.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, *123*(11), 2189–2202.

Broekens, J., Jacobs, E., & Jonker, C. M. (2015). A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, *27*(3), 215-233.

Dalgleish, T. (2004). Timeline: The emotional brain. *Nature Reviews Neuroscience*, *5*(7), 583–589.

Damasio, A. R. (1998). The somatic marker hypothesis and the possible functions of the prefrontal cortex. In A. C. Roberts, T. W. Robbins, & L. Weiskrantz (Eds.), *The prefrontal cortex: Executive and cognitive functions.* Oxford: Oxford University Press.

Damasio, A. R. (2006). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Vintage.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, *30*(2), 239–271.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain*, *126*(8), 1830–1837.

Fellows, L. K., & Farah, M. J. (2005, 2004). Different Underlying Impairments in Decision-Making Following Ventromedial and Dorsolateral Frontal Lobe Damage in Humans. *Cerebral Cortex*, *15*(1), 58–63.

Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., & Polkey, C. E. (2003). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain*, *126*(7), 1691–1712.

Krawczyk, D. C. (2002). Contributions of the prefrontal cortex to the neural basis of human decision making. *Neuroscience and Biobehavioral Reviews*, *26*(6), 631–664.

Kuleshov, V., & Precup, D. (2000). Algorithms for Multi-Armed Bandit Problems. *Journal of Machine Learning Research*, *1*, 1–48.

Maia, T. V., & McClelland, J. L. (2005). The somatic marker hypothesis: still many questions but no answers. *Trends in Cognitive Sciences*, *9*(4), 162–164.

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Moerland, T. M., Broekens, J., & Jonker, C. M. (2018; 2017). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, *107*(2), 443–480.

Osmankovic, D., & Konjicija, S. (2011). Implementation of Q - Learning algorithm for solving maze problem. In *Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1619–1622). USA: IEEE.

Plutchik, R. (2003). *Emotions and Life : Perspectives from Psychology, Biology, and Evolution*. Washington, DC: American Psychological Association.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (2nd ed.). Hoboken, N.J: Wiley.

Puviani, L., & Rama, S. (2016). A System Computational Model of Implicit Emotional Learning. *Frontiers in Computational Neuroscience*, *25*(56), 1–26.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), . New York: Appleton-Century-Crofts.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

Rolls, E. T. (2000). The Orbitofrontal Cortex and Reward. *Cerebral Cortex*, *10*(3), 284–294.

Rolls, E. T. (2013). *Emotion and Decision-Making Explained* (First ed.). Oxford New York, NY: Oxford University Press.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. New York, New York: Academic Press, Inc.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. Chichester; Hoboken, N.J: Wiley-Interscience.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second ed.). London; Cambridge, Massaschusetts: MIT Press.

Volz, K. G., & Hertwig, R. (2016). Emotions and Decisions: Beyond Conceptual Vagueness and the Rationality Muddle. *Perspectives on Psychological Science*, *11*(1), 101–116.

Wallis, J. D. (2007). Orbitofrontal Cortex and Its Contribution to Decision-Making. *Annual Review of Neuroscience*, *30*(1), 31–56.

Watkins, C. (1989). *Learning From Delayed Rewards*. Doctoral dissertation, King's College, London, UK.