Accepted Manuscript

Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality

Dimitri Ognibene, Vincenzo G. Fiore, Xiaosi Gu

PII: S0893-6080(19)30128-5
DOI: https://doi.org/10.1016/j.neunet.2019.04.022
Reference: NN 4152

To appear in: *Neural Networks*

Received date : 7 November 2018
Revised date : 6 April 2019
Accepted date : 25 April 2019

Please cite this article as: D. Ognibene, V.G. Fiore and X. Gu, Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. *Neural Networks* (2019), https://doi.org/10.1016/j.neunet.2019.04.022

Addiction beyond pharmacological effects: the role of environment complexity and bounded rationality

Dimitri Ognibene[1,2]

[1] School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

[2] ETIC, Universitat Pompeu Fabra, Barcelona, Spain

Vincenzo G. Fiore

Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Xiaosi Gu[a,b,c]

[a] Department of Psychiatry, [b] Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[c] The Mental Illness Research, Education, and Clinical Center (MIRECC VISN 2) at the James J. Peter Veterans Affairs Medical Center, Bronx, NY

Abstract

Several decision-making vulnerabilities have been identified as underlying causes for addictive behaviours, or the repeated execution of stereotyped actions despite their adverse consequences. These vulnerabilities are mostly associated with brain alterations caused by the consumption of substances of abuse. However, addiction can also happen in the absence of a pharmacological component, such as seen in pathological gambling and videogaming. We use a new reinforcement learning model to highlight a previously neglected vulnerability that we suggest interacts with those already identified, whilst playing a prominent role in non-pharmacological forms of addiction. Specifically, we show that a dual-learning system (i.e. combining model-based and model-free) can be vulnerable to highly rewarding, but suboptimal actions, that are followed by a complex ramification of stochastic adverse effects. This phenomenon is caused by the overload of the capabilities of an agent, as time and cognitive resources required for exploration, deliberation, situation recognition, and habit formation, all increase as a function of the depth and richness of detail of an environment. Furthermore, the cognitive overload can be aggravated due to alterations (e.g.

[1] Corresponding author dimitri.ognibene@gmail.com

[2] Dr Dimitri Ognibene has been supported by the VolkswagenStiftung project COURAGE Ref: 95564, H2020 FETPROACT-01-2018 project POTION ID: 824153, and FP7-COFUND 600387. Dr Xiaosi Gu is supported by NIDA R01DA043695 and the Mental Illness Research, Education, and Clinical Center (MIRECC VISN 2) at the James J. Peter Veterans Affairs Medical Center, Bronx, NY.

caused by stress) in the bounded rationality, i.e. the limited amount of resources available for the model-based component, in turn increasing the agent's chances to develop or maintain addictive behaviours. Our study demonstrates that, independent of drug consumption, addictive behaviours can arise in the interaction between the environmental complexity and the biologically finite resources available to explore and represent it.

Keywords: addiction, reinforcement learning, computational psychiatry, gambling, internet gaming, bounded rationality, exploration-exploitation;

Introduction

Addiction is marked by the compulsive execution of stereotyped actions despite their adverse consequences [1, 2, 3, 4, 5]. This maladaptive form of decision making is typically associated with the consumption of substances of abuse, such as alcohol, tobacco, illicit and prescription drugs [3, 6, 7]. More recently, the definition has been also used to describe gambling [8, 9] and other putative forms of behavioural addictions, such as internet gaming [10]. Importantly, these latter forms of addiction lack the neuro-pharmacological effects of a consumed drug, and yet are characterised by a striking similar symptomatology.

Several theories and computational models have been proposed to explain the repetition of suboptimal decisions typical of addiction [9, 3, 6, 7]. These theories assume decision making results from the interaction of multiple systems, e.g. habitual, deliberative, Pavlovian, motivational, situation identification, etc. which rely on different learning and computing principles. This composed structure is associated with several vulnerabilities to the pharmacological effects of drugs of abuse, each of which can result in the expression of compulsive repetition of drug intake [3, 11, 12, 13, 14, 15].

In particular, Reinforcement Learning (RL) models of addiction frequently assume that aberrant drug-seeking habits come to dominate behaviour in addiction due to drug induced biochemical hijacking of the dopaminergic prediction error signal [7, 16, 3, 17, 18, 19, 20]. The hypothesis of the dominance of the habitual system nicely accounts for aspects of addiction such as inelastic behaviour in the face of changes in the environment or even in presence of punishing outcomes following drug consumption [16, 21]. However, several other behaviours associated with addiction are left unaccounted for [18, 3]. First, one of the defining characteristics of substance abuse according to the DSM-5 is "A great deal of time is spent in activities necessary to obtain the substance (e.g., visiting multiple doctors or driving long distances)" [22]. Such temporally extended activities are often novel, complex and context-dependent [23, 18, 3, 24], and therefore are not driven by habitual processes or stimulus-response conditioning. Second, phenomena such as craving can occur even without exposure to conditioned stimuli (but see [25, 26]). Finally, gambling [8, 2, 1] and internet gaming [10], which are also considered part of the addictive behaviours, lack the pharmacological interference that is considered essential to drive the aberrant habit formation [9].

These issues have been partially addressed by hypothesising the presence of vulnerabilities affecting the deliberative system [3]. In particular, it has been suggested that non-habitual forms of addictive behaviours may be caused by errors of interpretation, where either the outcome of an action (drug consumption, gambling etc.) is over-evaluated as beneficial or useful or the long term consequences of these actions are under-evaluated in their negative effects. However, the computational mechanisms by which both drug-related and non drug-related addiction can induce these effects on the deliberative/planning system are not well understood [11, 27, 28].

Other models [9] have posed that addiction can emerge in environments characterised by incomplete or inaccessible information. Under these conditions, the underestimation of the negative consequences or the over-evaluation of the positive ones is simply caused by a lack of information. However, this hypothesis does not seem to match with clinical evidence, as once the required information is made readily available to addicted individuals, motivating their abstinence, relapse should not occur.

We propose a solution can be found in the analysis of the discrepancy between the resources available to an agent and those required to explore, represent or compute the environment it operates in. Most computational models of addiction have so far focused on environments characterised by the presence of easy to compute outcomes, where the number of actions available and their ramifications were limited. This simplification has distanced the computational analysis from the clinical practice which has long considered a wide range of environmental factors, and social interactions in particular, to have a strong impact on addiction development and maintenance [29, 30, 31].

Environment complexity and exploration are well recognised factors in the fields of Artificial Intelligence (AI) [32, 33], as well as developmental and computational neuroscience, in particular when considering the problem of the exploration-exploitation trade-off [34, 35, 36, 37, 38]. As the amount of experience required by an agent to achieve a specific behavioural performance grows faster than the product of the number of available states and actions [39, Chapter 8], exploration and training in complex environments can easily result in incomplete or incorrect representations of action-outcome ramifications [37, 40, 41]. Furthermore, if a complex environment is correctly represented in the agent's internal model, e.g. after a prolonged exploration, the stored action-outcome ramifications might still overload the agent's capacity to *internally assess* its available options. This inherent inadequacy of resources can be also aggravated by temporary forms of cognitive impairments which would dynamically increase the chances to trigger suboptimal planning. Interestingly, anxiety or stress are good examples of dynamic processes associated with temporary cognitive impairments and represent known triggers in addiction disorders and relapse after treatment [42, 43, 44, 45].

Our simulations show that the development of addictive behaviours may be supported by the interaction between specific features of the environment and both habitual and deliberative processes [37, 40, 46, 47]. We propose this vulnerability complements and interacts with previously described ones, capturing the emergence of addiction in the absence of pharmacological factors.

Materials and Methods

Agent. The behaviour of our simulated agents (Fig 1) is controlled by a hybrid (or dual) RL model system [48, 18, 49, 50, 51]. This algorithm maximises expected cumulative rewards by simultaneously learning, through a model-free (MF) component, and computing, through a model-based (MB) component, an optimal action strategy, or policy π.

The MF component is implemented as a standard tabular Q-learning algorithm [52]. MF algorithms such as Q-Learning and actor-critic architectures [53] are usually employed to model habitual behaviours [54, 55, 3, 48, 56, 57] and therefore are tipically associated with the dorsal cortico-striatal neural circuit [49, 58]. These algorithms are characterised by limited flexibility but computational efficiency as they require limited resources to slowly update associations between state-action pairs and values $\tilde{Q}^{MF}(s, a)$, depending on experience. The MB algorithm is employed to implement planning processes [11, 27, 26, 41] on the basis of an explicit representation, in an *internal model*, of action-state relationships and associated rewards, as experienced in the environment. Due to the similarity with goal-oriented processes, the MB component is often associated with the ventral cortico-striatal circuit [49, 58]. Where the MF component simply selects the best action among those available in its current state, the internal model of action-state sequences allows the MB component to evaluate entire policies, as if navigating decision trees with their ramifications and consequences, before making any decision. Such a process of evaluation is demanding in terms of computational resources and time, but allows a high degree of flexibility.

Most dual models assume an ideal MB process [50, 59], characterised by a complete knowledge of the environment and unlimited computational resources, which therefore always leads towards optimal choices. However, biological MB system are constrained, or *bounded*, by their limited resources [60, 61, 62, 63, 64, 65, 66, 67]. Thus, to model biologically plausible healthy and dysfunctional behaviours (as e.g. in addiction [18, 3]), in our simulations we have employed a MB component that represents only direct experience, and that relies on bounded computational resources [60, 61] to navigate its internal model. Importantly, our MB component generates a new value estimation at each step by applying the Bellman Equation a limited number of times to states sampled stochastically, following an early-interrupted variation of the Prioritized Sweeping algorithm [68], with stochastic selection of the states to update (see Algorithm 1). This is similar to what is regularly done in the Monte-Carlo Tree Search family of algorithms [69], which is commonly adopted in Artificial Intelligence for complex environments models where estimations over simulations are easier than complete bellman backups. However, the Early Interrupted Stochastic Prioritized Sweeping algorithm employed here is computationally more efficient for small environments [70], so to provide stable results with a limited number of updates.

Result: Q values
initialization;
$\forall s \quad H(s) = 0, V(s) = 0;$
steps=0 ;
while *steps* $< N_{ps}$ **do**
  steps=steps+1;
  $\tilde{s} \sim \eta \exp(\frac{H(s)}{T_{MB}})$`// sample state to update with softmax of H;`
  $\forall a \quad Q(\tilde{s}, a) = \sum_{s'} p(s'|\tilde{s}, a) \, [R(\tilde{s}, a, s') + V(s')];$
  $M = \max_a Q(\tilde{s}, a);$
  $\Delta = |V(\tilde{s}) - M|;$
  $V(\tilde{s}) = M;$
  $\forall s \quad h(s) = \Delta \times \max_a P(\tilde{s}|s, a);$
  $H(\tilde{s}) = h(\tilde{s});$
  $\forall s \neq \tilde{s} \quad H(s) = \max(h(s), H(s));$
end

Algorithm 1: Early Interrupted Stochastic Prioritized Sweeping pseudocode

In keeping with existing literature [11], we assumed that the MB and MF components do not share a common representations, and they do not interact during the computation of the respective state action values. However, a hybrid value function Q^{MX} is computed by balancing MF ($Q^{MF}(s, a)$) and MB (Q^{MB}) estimates depending on a parameter, β, as follows:

$$Q^{MX}(s, a) = \beta Q^{MB}(s, a) + (1 - \beta) Q^{MF}(s, a) \tag{1}$$

Similar to a previous study [58], six values (1, 0.8, 0.6, 0.4, 0.2, 0) have been used for this parameter to simulate different behavioural phenotypes, along a spectrum between purely model-based (β=1) and purely model-free (β=0) reinforcement learning. In terms of neural implementation, these phenotypes loosely match the neural systems dominated by either a ventral or a dorsal cortico-striatal circuit, with the strength of the directed connectivity between these circuits as the analogue of the beta values in the algorithmic model.

Finally, the agents selected the actions that were expected to maximize the future utility (Q^{MX}) in 90% of their selections. For the remaining 10% of selections, the agents would perform a random action, in a standard strategy meant to preserve exploration for all stages of the simulations, termed stochastic ϵ-greedy *persistent exploration* [71].

Environment. We tested our hypothesis that suboptimal, addiction-like, behaviours can emerge without pharmacological interference or MB-MF malfunctions, in an environment (Fig 2) that allows long action-sequences characterised by deep ramifications. In comparison with simpler environments, characterised by limited interactions or depth of action sequences (e.g. an operant conditioning chamber), environments simulating open space navigations require larger amount of resources invested in the exploration and computation of the action-outcome contingencies. Thus, the agents struggle to find and pursue those policies that lead to reward maximization (i.e. optimal behaviours) and to avoid those policies that lead overall losses (i.e. suboptimal behaviours).

Importantly, we could not investigate the same phenomena by including, for instance, a high discount factor in a simplified environment, as there are fundamental differences between disregarding temporally distant events and failing in exploring, representing and evaluating them. In fact, with a high discount, an addictive behaviour that disregards long term negative effects would be formally optimal and therefore it would not induce that sense of inability to stop [19] that often characterizes addiction.

The simulated agents operated under two different configurations of the environment or phases (Fig 2). Under the initial *safe phase* ($d_{init} = 50$ steps), the agents could only experience a moderate reward (termed *healthy* reward, $R_g = 1$) if they accessed the relative state. Once the healthy reward state was reached, an agent would be brought back to the initial state and could pursue the reward again. No other reward or punishment was available in any other part of the environment. Under the second *addiction phase* ($d_{drug} = 1000$ steps) the agent was still rewarded by accessing the healthy reward state, but it could also access a state characterised by a high reward (termed *addictive* reward, $R_d = 10$). This state was inescapably followed by a more unpredictable and mixed-in-value negative *after-effect segment* of the environment, which ideally simulated the multifaceted effects addictive behaviour has on the social life and health of the addicted individual. At the end of this after-effect segment, the agent would be again brought back to the initial state. Table 1 shows the number of updates that the original Prioritized Sweeping algorithm would have used to find the optimal policy in each phase. These are two orders of magnitude larger than the updates allowed by the adopted bounded MB.

Finally, to test the ability of the agents to adapt to changes, we modified the environment structure in a separate set of simulations. This modified environment included three arms in a Y shape, adding a segment to the two already described. This third segment -termed *neutral*- was kept empty, and reaching its end did not send the agent back to the starting position (as for the healthy reward state) or have it enter an after-effect segment (as for the addictive reward state), but it allowed the agent to freely move to the adjacent neutral states. After the time step 2500, the healthy reward (and its associated rule of sending back the agent to the origin point) was moved from its initial position to the end of the neutral segment. At the same time, the healthy reward segment became neutral (i.e. deprived of any reward), also inheriting the rule of free state transitions among neutral states instead of leading back to the initial state.

Phase	Number of Updates
Init	4,712
Addictive Reward	5,005

Table 1: Number of updates necessary to Prioritized Sweeping to find the value function for each phase

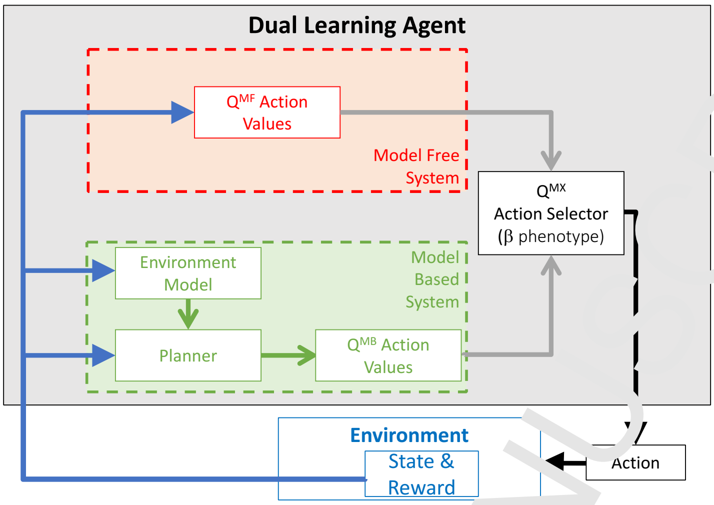

Figure 1: **Dual Learning Agent.** The decision making architecture includes: (i) a model free component (MF), which updates action values using value prediction error computations; and (ii) a model based component (MB), which generates an internal model of the environment, based on experienced action-outcomes and bounded computations. Action-outcome estimations derived from the two components are combined linearly according to a balance parameter, β, to drive action selection.

Table 2: **Environment Model Parameters**

Name	Description	Value
N_T	Number of states	22
N_G	Number Goal States	1
N_D	Number Addictive Area States	15
N_n	Number Neutral States	6
N_a	Number of actions	9
S_0	Starting state	4
R_p	Punishment end of Addictive Area	-4
R_c	Punishment in Addictive Area	-1.2
R_{dd}	Reward at entering Addictive reward state	10
R_g	Reward when entering healthy reward state	1
d_{init}	Duration safe phase	50
d_{drug1}	Duration addictive phase	1000

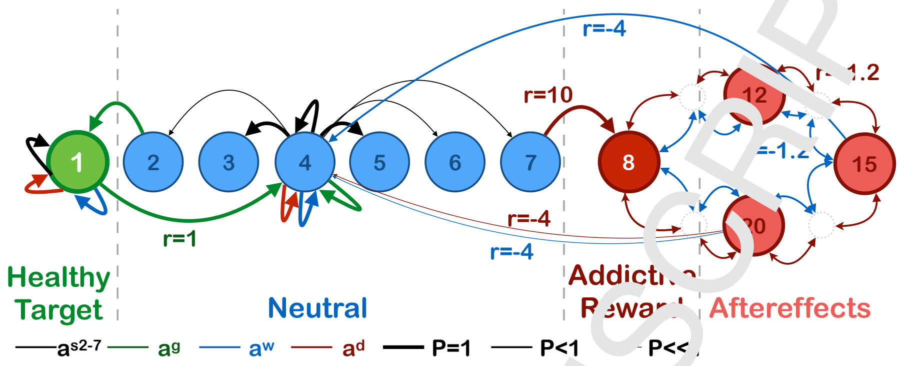

Figure 2: **Illustrative representation of the environment.** The states are disposed in a linear arrangement: on the left (number 1) a state associated with a healthy reward, on the right (number 8) a state associated with an addictive non-drug reward (e.g. gambling), separated by 6 neutral states that can be freely traversed. Entering the healthy reward state results in a moderate reward ($R_g = 1$), after which the agent returns to the central neutral state (number 4). Entering the addictive state provides an immediate high reward ($R_d = 10$), followed by a further segment of 14 states that are associated with negative outcomes (-1.2) or punishments. Within this segment of *after-effects* action results are stochastic, making it difficult for the agent to find a way out of this part of the environment, and resulting in an average overall punishment that makes the selection of the addictive reward suboptimal. In this illustrative representation, few key transitions are reported, with detailed descriptions for the states 1,4,15 and 20 for which line width represents transition probabilities and colour represents the action class (a_s, a_g, a_w, a_d). Neutral states can be crossed by selecting actions a_{s2-7}, which are deterministic for adjacent state while have high chance of failing for distant states. Agents can reach the healthy reward state by executing action a_g whilst in state 2, and the addictive reward state, by executing action a_d whilst in state 7. In the after-effect segment, actions results are less predictable and only action a_w at state 15 has a high chance of leaving the addictive area, with a cost of -4. All details about the environment are reported in table 2.

Table 3: **Agent Model Parameters**

Name	Description	Value
α	MF learning factor	0.05
γ	Discount factor	0.9
d_{MB}	MB decay factor	0.01
MBUS	Number of MB updates	50
T_{MB}	Temperature for stochastic state update selection	1
ϵ	Exploration Factor	0.1

Results

Independent of differences in the parametrisations regulating MB/MF balance, agents seem to rapidly acquire a stable behaviour, marked by the near-exclusive preference for either the healthy or the addictive state (Fig 3). This bifurcation into either an optimal (healthy) or a suboptimal (addictive) be-

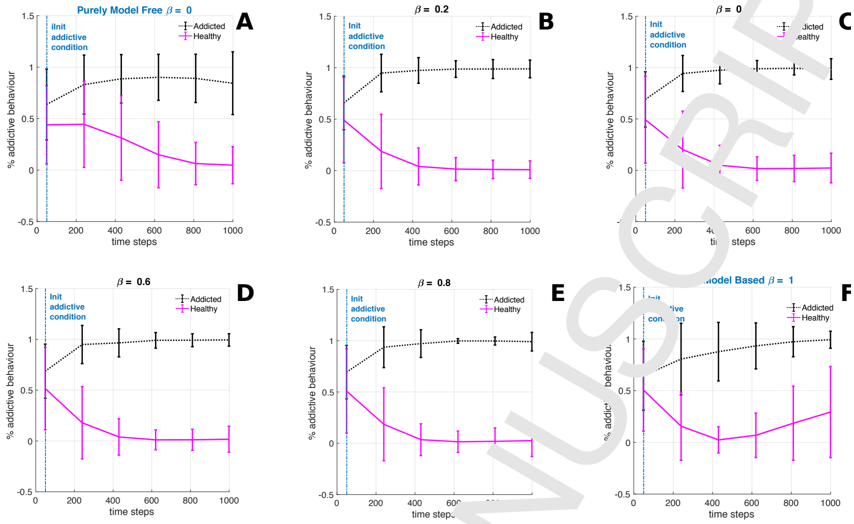

Figure 3: **Behavioural trajectories illustrating the ratio of healthy vs addictive reward state selections for addicted and healthy subjects.** The six panels highlight different behavioural trajectories, depending on β values, which represent MB/MF balance per population. Addicted agents are defined as those visiting the addictive reward state (number 8) more often than the healthy reward state for the whole experiment duration (0:1000 steps). Healthy subjects are defined by subtraction. Each of the six configurations of β values was tested with a total of 900 agents (healthy+addicted). Each data point in the chart reports mean and standard deviation for the number of visits to either the addictive of the healthy reward state, over the sum of the total visits to either state, across the 900 agents. A bifurcation in choice preference clearly emerges between addicted and healthy agents, for all parametrizations.

haviour trajectories is determined by few initial choices. The *healthy behaviour* is reached after less than 300 steps, across populations, and it is maintained for the entire time-length of the experiment. Conversely, the *addiction trajectory* is characterised by long-lasting, albeit transient, choice preferences, which are reached after less than 100 steps. Long simulations employing agents controlled uniquely by the MF component have proven the length of this transient stability is significant. These agents converge to optimum after around 100k steps (Fig 4), in comparison with the 300 steps required by the healthy agents, with identical parametrization, to engage in the optimal behaviour (cf. [52]). It must be noted that the MF component is a standard Q-Learning agent which has been formally proved to converge and which can be easily used to reproduce previous findings related to addiction, once the algorithm is used in association with easy to explore and compute environments [16].

In a previous study (cf. [58]), we demonstrated across algorithmic and neural implementation that the balance between MB and MF components significantly affected the chances to develop addictive behaviours, as higher resistance to addiction was found in populations characterised by *intermediate* values of β (Fig 5).

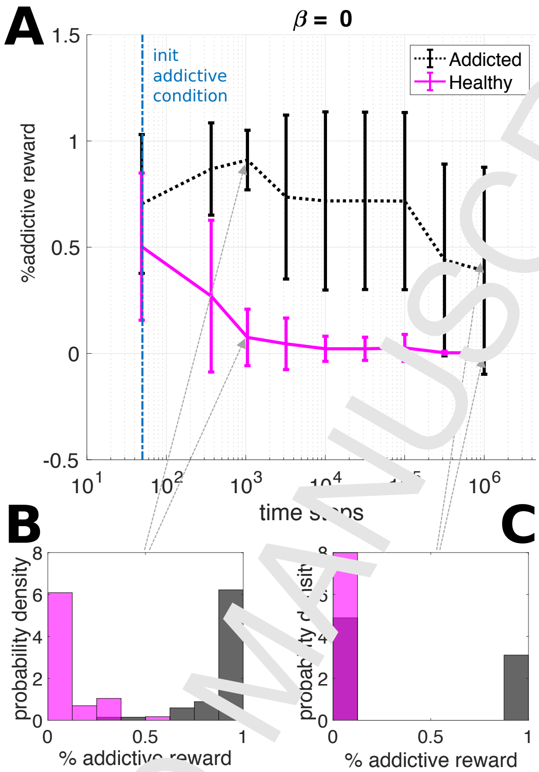

Figure 4: **Long runs with logarithmic time scale.** Behaviour expressed by purely MF agents ($\beta = 0$) was recorded and averaged over 100 runs, separating the addicted agents (high preference for the addictive reward state in the first 10,000 steps) from the healthy ones (the remaining agents, which showed the opposite preference within the same time period). A clear bifurcation emerged in the behaviour of the agents (cf. panel A with Fig. 3). Most of the addicted subjects changed their policy towards a healthy behaviour within a time of 200K steps. Histograms in panels B and C also illustrate the behavioural bifurcation, as the behaviour falls either in the interval with the lowest drug intake preference (0-0.125) or in the interval of the highest intake (0.875-1).

We further investigated these changes in the addiction development probabilities, using the amount of the available cognitive resources as a new independent dimension. The amount of these resources directly determines the depth of navigation in the internal model and, indirectly, how accurately such model is generated. Therefore, limited resources result in incorrect representation and action-sequence assessments, leading to suboptimal choices. To converge to optimum, when the model of the environment is known, the prioritized sweeping algorithm used in the MB requires above 4K updates of the value function. Note that for these internal *iterations steps*, the value of reaching a state is estimated using the internal model (fixed) without any actual interaction with the world

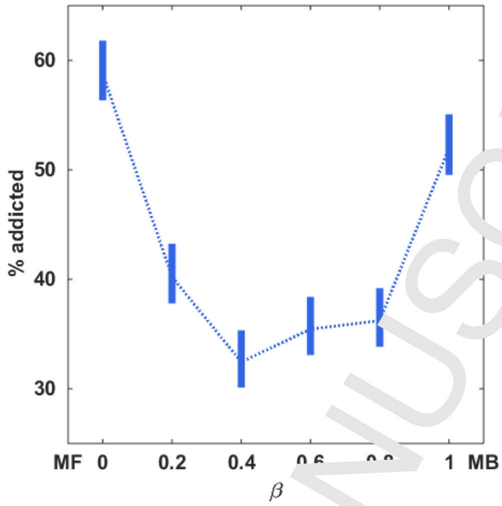

Figure 5: **Percentage and confidence intervals of addicted subjects per population, varying β.** Different β values controlling the balance between MF and MB components were used for distinct populations of 900 simulated subjects. Addicted agents are those that during the observation period, 1000 time steps, acquire the addictive reward more often than healthy reward. The percentage of addicted agents per population varied as a function of β values, where intermediate values showed a lower percentage of addicted agents (cf. [58]). Confidence interval were estimated assuming two-tail distribution and 95% confidence.

(Table1). Fig 6 shows that the chances to pursue suboptimal behaviours, i.e. seeking the addictive reward state, are inversely correlated with the resources available for the MB component (which we tested in a range well below the 4K updates necessary for optimal estimation). For instance, the population bounded by 50 Model Based Updates per Step (MBUS) resulted in 50% of subjects expressing addiction-like behaviours after 1K time steps, rising up to 90% of the subjects, after 10K time steps. At the opposite side of the spectrum, populations characterised by high computational resources (e.g. the tested 500 MBUS population) resulted in up to 20% of addicted subjects at 1K time steps, but this percentage falls to 0%, after 10K time steps, showing the agents had developed a correct model of the environment by that moment in the simulation. Contrarily to the MB-MF balance dimension, the behavioural trajectories caused by changes in the available cognitive resources are meaningful only when considered jointly, or in interaction, with the environment complexity. Any increase in the degree of complexity for the environment results in an increased demand of resources, to keep constant the likelihood of convergence to optimum. Ecological environments, however, are not limited by the artificial constraints of a laboratory or simulation set-up, so that they may require prohibitive and biologically implausible amounts of resources and exploration to replicate a result close to the described 500 MBUS population trajectory (see [39, Chapter 8] for

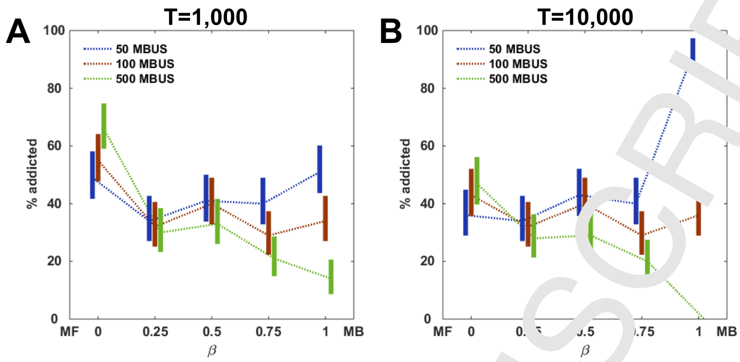

Figure 6: **Preference ratios and confidence intervals of agents expressing addiction-like behaviour within each parametrization of cognitive resource bounds (Model Based Updates per Step [MBUS]) and MB/MF balance factor β.** Initial performance (panel A, analysis on the behaviours in the interval 900 to 1000 timesteps) shows a significant preference for the selection of the addictive reward state across all values of β and most bounds for cognitive resource, with a low for very high resources (500MBUS), in association with $\beta = 1$. Towards the end of the simulation (panel B, interval 9900 to 10000 timesteps), we found that the populations diverge depending on the amount of cognitive resources available, as preference for the addictive state disappeared in the population characterised by very high resources and $\beta = 1$. Balanced MB-MF parametrizations (intermediate β values) were found generally more resistant to addiction, across values of cognitive bounds. A comparison between panels A and B illustrates the effects of exploration across all the parametrizations. Low values of β, dominated by the MF component, slightly reduce the number of addicted subjects after the first 10K steps, for all levels of cognitive resources, as the number of addicted agents remains above one third of the entire population. Exploration and experience with high values of β has opposite results, depending on the available cognitive resources. High cognitive resources, jointly with long exploration, lead to a strong reduction of addicted agents, suggesting a correct internal model of the environment is achieved through experience. With low cognitive resources, jointly with a strong MB component (high β), experience brings a substantial increase in the number of addicted agents. This result is due to a combination of poor environment representations and limited planning capabilities. Confidence intervals were estimated assuming a two-tail distribution and 95% confidence, with 100 simulated subjects per β value.

related theoretical proofs and [72] for experimental results with state-of-the-art supercomputers over more complex but still simplified environments).

We hypothesised that the observed behavioural bifurcation, i.e. the diverging behaviours displayed by two identical simulated agents placed in the same environment, was caused by the stochastic nature of the initial exploration phase. We assumed that during this phase, limited knowledge of the environment for both MF and MB components led to non-informative Q-values (i.e. the action-outcome estimations) and therefore to the execution of stochastic action selections. In turn, these initial choices determined which part of the environment would be explored and which would be neglected, shaping the value estimations and further biasing future exploration (cf. [9]).

To test this hypothesis we exposed our agents to the preliminary suboptimal-reward-free simplified environment for a longer time, thus granting early acqui-

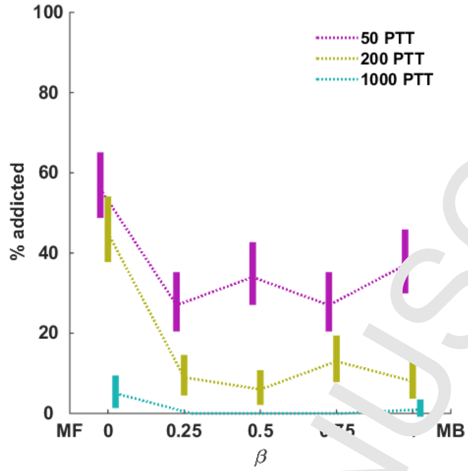

Figure 7: **Changes in behavioural trajectories as a function of pre-training time (PTT, timesteps in safe phase) and β parameter (MB/MF balance).** Exposure to the environment before the introduction of the addictive reward decreased the probability of addiction across all sets of parameters or populations. Extreme values for the parameter regulating MB/MF balance (i.e. $\beta \in \{0, 1\}$) resulted in a residual tendency to addiction even with long exposure. The chart reports confidence intervals for populations tested for 10K steps and composed by 100 agents under each condition, with an evaluation of the behavioural choice selections on the last 1K steps. Confidence intervals were estimated assuming a two-tail distribution and 95% confidence.

sition of an healthy action policy (Fig 7). Under this condition, the agents explored the environment before the introduction of the addictive reward, for a pre-training time (PTT), which lasted a variable number of time steps (50, 200 and 1000). Higher PTT were associated with a better representation of the policy required to reach the healthy state. However, the use of a constant exploration (ϵ-greedy) forced the agents to occasionally reach the addictive state reward, after it was introduced in the environment. Despite these exposures to the addictive reward, the chances to develop addiction after a PTT substantially decreased (Fig 7)across values of the parameter β, whilst confirming the general resistance to addiction of the balanced MB-MF systems (intermediate values of β).

Finally, we tested whether sudden environment changes could ignite addiction in agents that had developed the optimal healthy strategy [45, 42]. Our simulations in a Y-maze environment, characterised by the described healthy and addictive reward plus a neutral segment, allowed to test changes in behavioural trajectories after a sudden swap of reward and associated rules between the healthy reward and the neutral segment. This alteration in the environment, taking place after time step 2.5k, when a behavioural policy is consolidated, required the agents to rely again on exploration and learn a new goal directed strategy. The results showed that after this change in the environment, a sig-

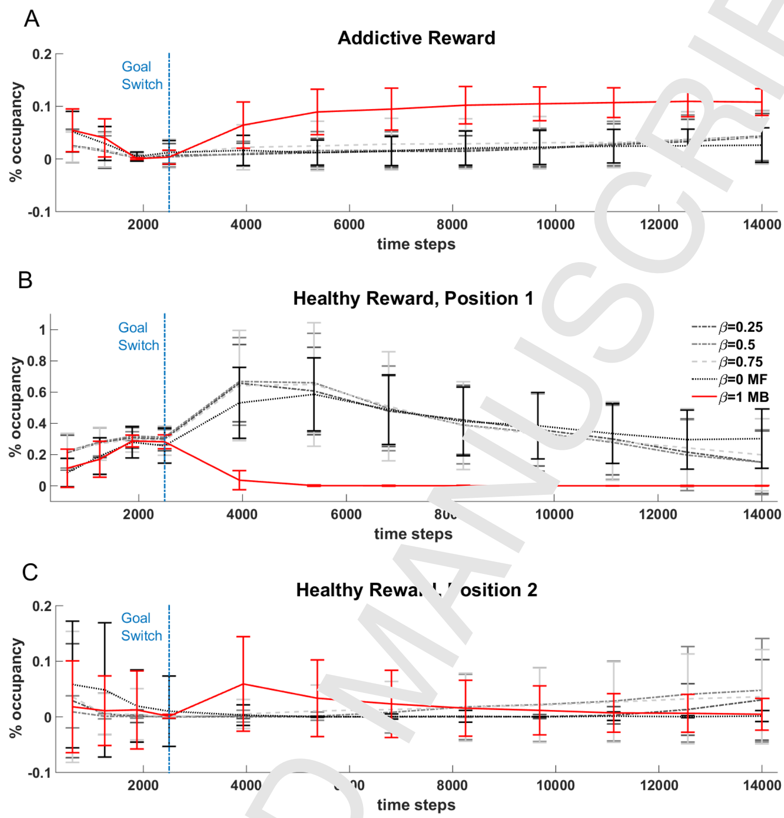

Figure 8: **Effects of environment change on healthy subjects** This figure illustrates the effects on behavioural trajectories caused by a change affecting the position of the healthy reward state, depending on the parameter β regulating MB-MF balance. The change takes place at time step 2500, when neutral and healthy reward segments are swapped while the addictive segment maintains its configuration. Purely model based agents ($\beta = 1$) switch rapidly to the addictive behaviour after the change, whereas agents with a non zero MF component gradually unlearn the acquired healthy policy to switch towards either the selection of the addictive state or the re-positioned healthy state. The increased number of visits for the first healthy reward position (panel B) is due to the sudden disappearance of the rewarded action that from this state used to lead (before the swap with the neutral segment) to the starting state (see Fig 2). Without this transition towards the starting state of the environment, the agent expresses cyclic exploratory behaviours, as it can re-enter the now neutral state as soon as it steps outside of it.

nificant portion of agents previously following a healthy policy developed a suboptimal addiction behaviour (Fig 8). Importantly, this test proved behavioural shifts to suboptimal behaviour could be induced by changes in the environment, in the absence of malfunctions of the decision components or any pharmacological interference.

Discussion

As formalised in a seminal work by Redish [16], the RL approach to addiction is based on the hypothesis that drug values are always underestimated by the MF learning system of a biological agent. This phenomenon is mediated by hyper-physiologic phasic responses of dopamine to drug consumption, which deceive the individual consuming the substance of abuse into perceiving the substance itself as always more rewarding than expected (i.e. a non compensable positive prediction error). In turn, this mismatch between expected and perceived outcomes results in an unlimited growth of the perceived value of drug related actions and aberrant reinforcement, causing habitual decision making, compulsive responses to drug-related stimuli and inelastic behaviour in the face adverse consequences [73, 74, 75, 12].

Despite significant advances in capturing important and complex features of addiction behaviour [19, 18, 11], this model remains primarily an expression of a malfunction of the MF component and therefore it leaves important questions unanswered [76, 20]. In particular, the role of the MB component in addiction is still unclear. First, even though interactions of deliberative computations with dopamine have been described [48], the effects of drug consumption on the generation and assessment of the internal representations of the environment have not been clarified. Second, phenomena such as craving, addiction behaviours which do not rely on stimulus-response habits (e.g. prolonged research for the preferred substance of abuse in novel environments), or non-pharmacological forms of addiction, all seem to suggest that the MB component plays a significant role in driving addiction-like suboptimal behaviours [11].

In this study, we have proposed that addiction-like behaviours can emerge in complex environments, if the dual-learning agent fails to correctly represent and compute action-outcomes associations, due to limited cognitive resources and exploration. In our simulations, a segment of the environment was designed so that an immediate high reward would be followed by multiple, inescapable and heavily stochastic, negative outcomes. We then tested different populations differing in the amount of available cognitive resources and found this variable was inversely correlated with the percentage of agents pursuing the addictive (sub-optimal) reward. Thus, stereotyped inelastic behaviours emerged in a fully accessible and explorable environment, despite the absence of a classic form of drug-induced aberrant prediction error signal or an otherwise malfunctioning MF system. This finding is consistent with previous studies indicating reduced contribution of the MB component may be a risk factor for addiction [77] and we argue it indicates a key computational process underlying those forms of addiction that are not based on the consumption of substances of abuse (e.g. gambling or videogaming).

Beyond the limitations of any experimental settings, the exponential growth of complexity that is associated with ecological environments could easily outmatch the equivalent growth of computational resources in a biologically plausible MB component. Furthermore, our results show that even purely deliberative agents with high cognitive capabilities can still be susceptible to addiction due to dynamic fluctuations in the exploration costs (i.e. sudden changes in the environment), or in the availability of computational resources (e.g. due to stress, a known trigger for addiction [78, 43, 66, 65, 79, 80, 61]). This ambivalency of the MB component in either protecting from or fosetering addiction depending on the amount of reseources it relies on, is consistent with multiple studies that have highlighted both decreased and increased neural responses in those brain areas associated with MB decision making, in addict individuals in comparison with controls. and depending on task and context [81, 82, 83, 84, 85].

This MB vulnerability can interact with previously described ones [3]. In forms of addiction dominated by the non-compensable prediction errors and hyper-physiologic DA responses, erroneous representations and assessments of the environment can aggravate the behavioural symptoms associated with the classic MF malfunction. This interaction can account for those complex non-habitual drug-seeking behaviours that are not triggered by the presence of drug-related stimuli [23, 18, 3, 24]. Importantly, a resource-bounded MB component may fail in evaluating long term action effects even after extensive exploration, so that even after the MF component has eventually converged towards an optimal behaviour (e.g. after a successful treatment), the MB component may keep pursuing sub-optimal policies, contributing to both craving and relapse [86]. Furthermore, by over-selecting the addictive reward early on in the task, exploration and representation of alternative routes in addicted agents remain limited, so that the stronger the addiction, the more compromised the model of the environment. This phenomenon, jointly with the fluctuations of long term outcome estimations under conditions of low MB resources [39, Sections 2.4-5], results in lowering the chances to disengage from pursuing the suboptimal policy at each step taken in the direction of the addictive reward, putatively simulating a context-related sense of inability to stop [19].

Finally, the vulnerability we have described can be seen as ideally contiguous with those associated with state identification errors [9, 87, 88, 89, 90]. Under conditions of the environment in which information about the states is either incomplete or inaccessible, the resulting interaction between state identification and value estimation can cause the creation of fictitious internal states, where addictive behaviours would always be considered as highly rewarding [9]. This hypothesis was originally proposed as a cause of context-driven addiction and has been used to describe gambling [9]. Under the conditions we have proposed, information exceeding an agent cognitive capabilities would be essentially lost to an agent, however the two vulnerabilities remain significantly different under many other aspects. The vulnerability we have described is not restricted to the opacity of a specific environment, and the dynamic interplay between exploration demands and availability of resources allowed us to account for the presence of different behavioural trajectories or phenotypes. We have observed that behavioural differences can arise from any change (either temporary or permanent) in the key parameter of the available cognitive resources, as well as unexpected changes in the environment structure or simply due to less than *few hundreds* initial stochastic exploration steps. These differentiations and behavioural trajectories took place despite the presence of a converging MF algorithm (as demonstrated in the *long run* tests) and it was neither caused by a disruption of the classical TD-MF learning mechanism [16, 19], nor by incomplete access to information concerning rewards and punishments in the environment [9].

Our findings have interesting implications for treatment development. A crucial problem is that the MB component is unlikely to increase its computational power with training, so that even if a correct model is formed, the agent might still pursue addictive behaviours, initiating relapse, due to difficulties in assessing complex ramifications associated with apparently rewarding initial choices. Thus, we hypothesise a treatment could aim at simplifying or making more explicit and accessible the structure of the environment. In doing so, normally occurring negative outcomes associated with the addictive behaviour would be easier to be taken into consideration and -importantly- courses of action leading to healthy policies would become competitive in the MB component. Unfortunately, there is the possibility that, independent of treatment, the MB component might keep associating a high reward to the addictive behaviour due to a stochastic representation of past experienced rewards, possibly modulated by reward intensity and distance in time. We hypothesise these conditions could be ameliorated by a conflict between MF and MB component, where addiction-avoiding habits could be developed during treatment, as suggested by our pre-training tests (Fig 7).

In conclusion, several studies focus on the effects that different sources of complexity (most prominently, social factors [91, 92] and stress [93, 45]) may have on addiction, however current computational modelling literature has often neglected these aspects [29, 31]. In this work we have proposed a step forward in the direction of more ecologically plausible simulations of healthy and dysfunctional behaviours, as we highlighted the interaction between limited MB resources and overwhelming representation requirements.

[1] American Psychiatric Association, Diagnostic and statistical manual of mental disorders: DSM-5, 5th ed. Edition, American Psychiatric Association, Washington, DC, 2013.

[2] M. Tschernegg, J. S. Crone, T. Eigenberger, P. Schwartenbeck, M. Fauth-Bhler, T. Lemnager, K. Mann, N. Thon, F. M. Wurst, M. Kronbichler, Abnormalities of functional brain networks in pathological gambling: a graph-theoretical approach., Frontiers in human neuroscience 7 (2013) 625. doi:10.3389/fnhum.2013.00625.

[3] A. D. Redish, S. Jensen, A. Johnson, A unified framework for addiction: vulnerabilities in the decision process, Behavioral and Brain Sciences 31 (04) (2008) 415–437.

[4] M. Griffiths, A components model of addiction within a biopsychosocial framework, Journal of Substance Use 10 (4) (2005) 191–197.

[5] M. Griffiths, Behavioural addiction: an issue for everybody?, Employee Councelling Today 8 (3) (1996) 19–25.

[6] P. Dayan, Dopamine, reinforcement learning, and addiction, Pharmacopsychiatry 42 (S 01) (2009) S56–S65.

[7] B. J. Everitt, T. W. Robbins, Drug addiction: updating actions to habits to compulsions ten years on, Annual Review of Psychology 67 (2016) 23–50.

[8] C. J. Rash, J. Weinstock, R. Van Patten, A review of gambling disorder and substance use disorders., Substance abuse and rehabilitation 7 (2016) 3–13. doi:10.2147/SAR.S83460.

[9] A. D. Redish, S. Jensen, A. Johnson, Z. Kurth-Nelson, Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling., Psychological review 114 (2007) 784–805. doi:10.1037/0033-295X.114.3.784.

[10] N. M. Petry, F. Rehbein, C.-H. Ko, C. P. O'Brien, Internet gaming disorder in the dsm-5., Current psychiatry reports 17 (2015) 72. doi:10.1007/s11920-015-0610-0.

[11] N. D. Daw, P. Dayan, The algorithmic anatomy of model-based evaluation, Philosophical Transactions of the Royal Society B: Biological Sciences 369 (1655) (2014) 20130478.

[12] J. P. ODoherty, S. W. Lee, D. McNamee, The structure of reinforcement-learning mechanisms in the human brain, Current Opinion in Behavioral Sciences 1 (2015) 94–100.

[13] D. A. Norman, T. Shallice, Attention to action: Willed and automatic control of behavior, Cognitive neuroscience: A reader (2000) 376–390.

[14] L. Nadel, Multiple memory systems: What and why, an update, Memory systems 1994 (1994) 39–63.

[15] N. J. Cohen, L. R. Squire, Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that, Science 210 (4466) (1980) 207–210.

[16] A. D. Redish, Addiction as a computational process gone awry, Science 306 (5703) (2004) 1944–1947.

[17] G. Di Chiara, Drug addiction as dopamine-dependent associative learning disorder, European journal of pharmacology 375 (1999) 13–30.

[18] D. A. Simon, N. D. Daw, Dual-system learning models and drugs of abuse, in: Computational Neuroscience of Drug Addiction, Springer, 2012, pp. 145–161.

[19] M. Keramati, B. Gutkin, Imbalanced decision hierarchy in addicts emerging from drug-hijacked dopamine spiraling circuit, PloS one 8 (4) (2013) e61489.

[20] A. Dezfouli, P. Piray, M. M. Keramati, H. Ekhtiari, C. Lucas, A. Mokri, A neurocomputational model for cocaine addiction, Neural computation 21 (10) (2009) 2869–2893.

[21] W. K. Bickel, L. A. Marsch, Toward a behavioral economic understanding of drug dependence: delay discounting processes, Addiction 96 (1) (2001) 73–86.

[22] National Insitute on Drug Abuse, The science of drug abuse and addiction: The basics.
URL `https://www.drugabuse.gov/publications/media-guide/science-drug-abuse-addiction-basics`

[23] S. T. Tiffany, A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes., Psychological review 97 (2) (1990) 147.

[24] D. H. Root, A. T. Fabbricatore, D. J. Barker, S. Ma, A. P. Pawlak, M. O. West, Evidence for habitual and goal-directed behavior following devaluation of cocaine: a multifaceted interpretation of relapse., PloS one 4 (2009) e7170. `doi:10.1371/journal.pone.0007170`.

[25] X. Gu, F. Filbey, A bayesian observer model of drug craving, JAMA psychiatry.

[26] A. D. Redish, A. Johnson, A computational model of craving and obsession., Annals of the New York Academy of Sciences 1104 (2007) 324–339. `doi:10.1196/annals.1390.014`.

[27] B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, N. D. Daw, Model-based choices involve prospective neural activity, Nature neuroscience 18 (5) (2015) 767–772.

[28] R. Kaplan, J. King, R. Koster, W. D. Penny, N. Burgess, K. J. Friston, The neural representation of prospective choice during spatial planning and decisions, PLOS Biology 15 (1) (2017) e1002588.

[29] M. Heilig, D. H. Epstein, M. A. Nader, Y. Shaham, Time to connect: bringing social context into addiction neuroscience., Nature reviews. Neuroscience 17 (2016) 592–599. `doi:10.1038/nrn.2016.67`.

[30] C. L. Hart, Viewing addiction as a brain disease promotes social injustice, Nature Human Behaviour 1 (2017) 0055.

[31] A. Reiter, A. Heinz, L. Deserno, Linking social context and addiction neuroscience: a computational psychiatry approach, Nature Reviews Neuroscience 18 (7) (2017) 450–450.

[32] S. J. Russell, P. Norvig, Artificial Intelligence A Modern Approach 3rd Ed., Prentice Hall, 2010.

[33] H. Geffner, B. Bonet, A concise introduction to models and methods for automated planning, Synthesis Lectures on Artificial Intelligence and Machine Learning 8 (1) (2013) 1–141.
URL `http://www.morganclaypool.com/doi/abs/10.2200/S00513ED1V01Y201306AIM022`

[34] M. Kearns, S. Singh, Near-optimal reinforcement learning in polynomial time, Machine Learning 49 (2-3) (2002) 209–232.

[35] A. L. Strehl, L. Li, M. L. Littman, Reinforcement learning in finite mdps: Pac analysis, The Journal of Machine Learning Research 10 (2009) 2413–2444.

[36] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, Journal of Machine Learning Research 11 (Apr) (2010) 1563–1600.

[37] T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, I. D. Couzin, C. S. R. Group, et al., Exploration versus exploitation in space, mind, and society, Trends in cognitive sciences 19 (1) (2015) 46–54.

[38] N. Cesa-Bianchi, C. Gentile, G. Lugosi G.and Neu, Boltzmann exploration done right, ArXiv Computer Science e-prints.

[39] S. M. Kakade, On the sample complexity of reinforcement learning, Ph.D. thesis, University of London (2003).

[40] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, G. Pezzulo, Active inference and epistemic value, Cognitive neuroscience 6 (2015) 1–28.

[41] M. Kearns, S. Singh, Finite sample convergence rates for q-learning and indirect algorithms, Advances in neural information processing systems (1999) 996–1002.

[42] R. Sinha, The role of stress in addiction relapse., Current psychiatry reports 9 (2007) 388–395.

[43] K. Starcke, M. Brand, Decision making under stress: a selective review., Neuroscience and biobehavioral reviews 36 (2012) 1228–1248. `doi:10.1016/j.neubiorev.2012.02.003`.

[44] S. Pabst, M. Brand, O. T. Wolf, Stress and decision making: a few minutes make all the difference., Behavioural brain research 250 (2013) 39–45. `doi:10.1016/j.bbr.2013.04.046`.

[45] J. R. Mantsch, D. A. Baker, D. Funk, A. D. Lê, Y. Shaham, Stress-induced reinstatement of drug seeking: 20 years of progress, Neuropsychopharmacology 41 (1) (2016) 335–356.

[46] P. Dayan, Exploration from generalization mediated by multiple controllers, in: Intrinsically motivated learning in natural and artificial systems, Springer, 2013, pp. 73–91.

[47] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, A. Baranes, Information-seeking, curiosity, and attention: computational and neural mechanisms, Trends in cognitive sciences 17 (11) (2013) 585–593.

[48] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors, Neuron 69 (6) (2011) 1204–1215.

[49] R. J. Dolan, P. Dayan, Goals and habits in the brain, Neuron 80 (2) (2013) 312–325.

[50] M. Keramati, A. Dezfouli, P. Piray, Speed/accuracy trade-off between the habitual and the goal-directed processes, PLoS computational biology 7 (5) (2011) e1002055.

[51] F. Cushman, A. Morris, Habitual control of goal selection in humans, Proceedings of the National Academy of Sciences 112 (45) (2015) 13817–13822.

[52] C. J. Watkins, P. Dayan, Q-learning, Machine learning 8 (3-4) (1992) 279–292.

[53] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, MIT PRESS, Cambridge, MA, 1998.

[54] K. Doya, Reinforcement learning in continuous time and space, Neural Computation 12 (1) (2000) 219–245.
URL `citeseer.ist.psu.edu/doya00reinforcement.html`

[55] N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nature neuroscience 8 (12) (2005) 1704–1711.

[56] A. Dezfouli, B. W. Balleine, Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized, PLoS Comput Biol 9 (12) (2013) e1003364.

[57] A. Dezfouli, B. W. Balleine, Habits, action sequences and reinforcement learning, European Journal of Neuroscience 35 (7) (2012) 1036–1051.

[58] V. G. Fiore, D. Ognibene, B. Adinoff, X. Gu, A multilevel computational characterization of endophenotypes in addiction, eNeuro.

[59] G. Pezzulo, M. A. van der Meer, C. S. Lansink, C. Pennartz, Internally generated sequences in learning and executing goal-directed behavior, Trends in cognitive sciences.
URL `http://www.sciencedirect.com/science/article/pii/S1364661314001570`

[60] H. A. Simon, Theories of bounded rationality, Decision and organization 1 (1) (1972) 161–176.

[61] S. Russell, E. Wefald, Decision Theoretic Control of Reasoning: General Theory and an Algorithm to Game Playing, Tech. rep. (1988).

[62] D. Ognibene, et al., Resources allocation in a bayesian, schema-based model of distributed action control., in: NIPS-Workshop on Probabilistic Approaches for Robotics and Control., 2009.

[63] Q. J. M. Huys, N. Eshel, E. O'Nions, L. Sheridan, P. Dayan, J. P. Roiser, Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees., PLoS Comput Biol 8 (3) (2012) e1002410. doi:10.1371/journal.pcbi.1002410.
URL http://dx.doi.org/10.1371/journal.pcbi.1002410

[64] K. Friston, P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens, R. J. Dolan, The anatomy of choice: dopamine and decision-making., Philosophical transactions of the Royal Society of London. Series B, Biological sciences 369. doi:10.1098/rstb.2013.0481.

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

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

[67] P. A. Ortega, D. A. Braun, Thermodynamics as a theory of decision-making with information-processing costs, Proc. R. Soc. A.

[68] A. W. Moore, C. G. Atkeson, Prioritized sweeping: Reinforcement learning with less data and less time, Machine Learning 13 (1) (1993) 103–130.

[69] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, et al., A survey of monte carlo tree search methods, Computational Intelligence and AI in Games, IEEE Transactions on 4 (1) (2012) 1–43.

[70] C. Domshlak, Z. Feldman, To uct, or not to uct?(position paper), in: Sixth Annual Symposium on Combinatorial Search, 2013.

[71] S. Singh, T. Jaakkola, M. L. Littman, C. Szepesvri, Convergence results for single-step on-policy reinforcement-learning algorithms, Machine Learning 38 (3) (2000) 287–308.
URL http://dx.doi.org/10.1023/A:1007678930559

[72] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354.

[73] H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal., Neuron 47 (2005) 129–141. doi:10.1016/j.neuron.2005.05.020.

[74] J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, R. J. Dolan, Temporal difference models and reward-related learning in the human brain, Neuron 38 (2) (2003) 329–337.

[75] S. M. McClure, G. S. Berns, P. R. Montague, Temporal prediction errors in a passive learning task activate human striatum, Neuron 38 (2003) 339–346.

[76] L. V. Panlilio, et al., Blocking of conditioning to a cocaine-paired stimulus: testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward, Pharm Biochem and Behav.

[77] A. M. Reiter, L. Deserno, T. Wilbertz, H.-J. Heinze, F. Schlagenhauf, Risk factors for addiction and their association with model-based behavioral control, Frontiers in behavioral neuroscience 10.

[78] H. Simon, Rational choice and the structure of the environment., Psychological review 63 (2) (1956) 129–138.

[79] S. Zilberstein, Metareasoning and bounded rationality, in: proceedings of the workshop on metareasoning of aaai 2008, 2008.

[80] S. J. Russell, Rationality and intelligence, Artificial Intelligence 94 (1-2) (1997) 57 – 77, economic Principles of Multi-Agent Systems. doi:DOI:10.1016/S0004-3702(97)00026-X. URL http://www.sciencedirect.com/science/article/B6TYF-3SP2BB3-3/2/6934[illegible]203efd0eeab1d905638092257a

[81] L. Nestor, R. Hester, H. Garavan, Increased ventral striatal bold activity during non-drug reward anticipation in cannabis users, Neuroimage 49 (1) (2010) 1133–1143.

[82] R. Z. Goldstein, N. D. Volkow, Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications, Nature Reviews Neuroscience 12 (11) (2011) 652–669.

[83] I. M. Balodis, M. N. Potenza, Anticipatory reward processing in addicted populations: a focus on the monetary incentive delay task, Biological psychiatry 77 (5) (2015) 434–444.

[84] D. W. Hommer, J. M. Bjork, J. M. Gilman, Imaging brain response to reward in addictive disorders, Annals of the New York Academy of Sciences 1216 (1) (2011) 50–61.

[85] E. H. Limbrick-Oldfield, R. J. van Holst, L. Clark, Fronto-striatal dysregulation in drug addiction and pathological gambling: consistent inconsistencies?, NeuroImage: Clinical 2 (2013) 385–393.

[86] J. Stewart, Psychological and neural mechanisms of relapse, Philosophical Transactions of the Royal Society of London B: Biological Sciences 363 (1507) (2008) 3147–3158.

[87] A. McCallum, Efficient exploration in reinforcement learning with hidden state, 1997.

[88] N. Tishby, D. Polani, Information theory of decisions and actions, in: Perception-Action Cycle, Springer, 2011, pp. 601–636. URL `http://link.springer.com/chapter/10.1007/978-1-4419-1452-1_19`

[89] F. Doshi-Velez, D. Pfau, F. Wood, N. Roy, Bayesian nonparametric methods for partially-observable reinforcement learning, Pattern Analysis and Machine Intelligence, IEEE Transactions on 37 (2) (2015) 394–407.

[90] D. Ognibene, N. C. Volpi, G. Pezzulo, G. Baldassare, Learning epistemic actions in model-free memory-free reinforcement learning: experiments with a neuro-robotic model, in: Biomimetic and Biohybrid Systems, Springer, 2013, pp. 191–203. URL `http://link.springer.com/chapter/10.1007/978-3-642-39802-5_17`

[91] L. F. Berkman, I. Kawachi, M. M. Glymour, Social epidemiology, Oxford University Press, 2014.

[92] B. E. Havassy, S. M. Hall, D. A. Wasserman, Social support and relapse: Commonalities among alcoholics, opiate users, and cigarette smokers, Addictive behaviors 16 (5) (1991) 235–246.

[93] R. Sinha, Y. Shaham, M. Heilig, Translational and reverse translational research on the role of stress in drug craving and relapse, Psychopharmacology 218 (1) (2011) 69