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

Dimitri Ognibene[1,2], Vincenzo G. Fiore[3], Xiaosi Gu[3]

1 School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
2 Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
3 School of Behavioral and Brain Sciences, University of Texas at Dallas, Dallas, TX, USA

*dimitri.ognibene@essex.ac.uk

Abstract

In this paper, we propose environment complexity, i.e. difficulty to find healthy rewards vs unhealthy ones and difficulty in computing adverse consequences following unhealthy rewards, may be a complementary, and mostly neglected, vulnerability in addiction. We show in a new RL model, that this vulnerability can significantly contribute to both heterogeneity of addictive behaviours and addiction in the absence of drug induced malfunctions of the standard RL decision systems. Multiple components and processes of RL decision systems are either directly vulnerable to environment complexity (e.g. deliberation, situation recognition) or indirectly affected, as exploration demands increase as a function of environment complexity (e.g. habits formation). We targeted these vulnerabilities in a computational experiment with dual-learning agents. We employed an environment where highly rewarding, but suboptimal, actions are followed by several undesired (punishing) effects with sparse peaks of intensity, in a pattern that challenges the computational capabilities of the agents. Our simulations show that the suboptimal decision to repeatedly pursue the high reward, despite its negative effects (i.e. addiction-like behaviour), emerges and is maintained for an extended period of time. This behaviour is found as a bifurcation in several individuals of uniform populations of dual-learning agents over a wide range of agent parametrisations. Our simulations offer a theoretical investigation into addiction arising independently of a pharmacological component (e.g. gambling and videogaming), highlighting the role played by specific conditions of interaction between environment complexity and the agent's computational capabilities.

Author Summary

This paper reports simulations based on RL 'dual-system theory of decision making'. They show that, in non-trivial environments, addiction-like behaviours can appear as a behavioural bifurcation in uniform populations of agents without any variation of the standard RL mechanisms. This result suggests that pharmacological effects of drugs or biochemical malfunctions of the decision system may not be the only factors leading to addictive behaviours. Addiction may also emerge as a result of the dynamic process of adaptation of an agent with limited computational capabilities to complex environments.

These computational findings highlight the importance of the interaction among environment, learning dynamics, and the computational capabilities of an agent developing addiction, suggesting these components should be taken into account when devising and modelling new treatments.

Introduction

Addiction is marked by the repeated execution of actions despite their adverse consequences [1–5]. 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]. Other pathological behaviours, such as compulsive buying [11], connected with addiction in folk psychology but missing the external biochemical factors have been discussed but finally not categorised as actual addiction in the DSM-5, notwithstanding the many behavioural similarities [12, 13].

Several theories and computational models have been proposed to explain the repetition of suboptimal decisions typical of addiction [3, 6, 7, 9]. In most of these theories, decision making is assumed to result from the interaction of multiple systems, e.g. habitual, deliberative, Pavlovian, motivational, situation identification, etc., relying on different learning and computing principles. This composed structure leads to complex interactions that have been seen as candidate causes of several vulnerabilities, which can be exploited by the pharmaceutical effects of drugs to trigger compulsive repetition of drug intake [3, 14–18].

Many current RL models of addiction assume that aberrant drug-seeking habits come to dominate behaviour in addiction due to drug induced bio-chemical hijacking of the dopaminergic prediction error signal [3, 7, 19–23]. This substance induced habitual dominance is particularly important to explain aspects of addiction such as reduced elasticity, i.e. the reduced ability to adapt to changes in the environment, as addictive behaviours are maintained even in presence of punishment [19, 24].

However, substance induced habitual dominance alone falls short in explaining several findings associated with addictive behaviours [3, 21]. 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)" [25]. Such temporally extended activities are often novel, complex and context-dependent [3, 21, 26, 27], and are unlikely to be justified by appealing solely on habitual processes or stimulus-response conditioning. In a similar manner, the hypothesis of substance induced habitual dominance also fails in explaining phenomena such as craving, which can occur even without exposure to conditioned stimuli (but see [28, 29]). Finally, not only different drugs present very different pharmacological effects, but addictive behaviours are described also for gambling [1, 2, 8] and internet gaming [10], which cannot rely on the hypothesis of pharmacological hijacking of stimulus-response habit generation [9].

These issues have been partially addressed by highlighting the presence of several vulnerabilities that are proposed to affect several components of the decision system [3]. Key vulnerabilities involving the deliberative system have been described as the (i) misinterpretion of a drug, or another outcome or action (shopping, gambling), as beneficial or useful for the agent; (ii) overevaluation of the *immediate* reward associated with drug assumption; (iii) overevaluation of the *sequences* of actions comprising drug intake and its consequences. However, the computational mechanisms by which both drug-related and non drug-related addiction can induce these effects on the planning system are not well understood [14, 30, 31].

In particular, even though interactions of deliberative computations with dopamine have been found, e.g. model based computations have been shown to participate in the formation of the dopaminergic prediction error supposedly used to train model free action selection [32], it is still debated whether dopamine can play a role in altering the learning of reward model in the planning system. Furthermore, it is not clear if such drug induced alteration of the deliberative system reward representation would resemble the alteration affecting habit formation. The deliberative system reward model depends on the current state, action and reward, whereas habit formation is achieved via alteration of value computation, which itself depends on several factors: the state reached, its value and, more importantly, the policy adopted [14, 21, 33–36]. In other words, dopamine appears more involved in representing 'secondary incentives or motivational values and not representations of one-step rewards' which is employed by the deliberative system [21, Section 5.3.1].

Regarding the state identification system, crucial for interacting in natural environment pervaded of missing information, an interesting vulnerability that does not require a pharmacological interaction [37–41] has been modelled in the RL framework by Redish and collaborators [9]. They showed that a putative interaction between state identification and value estimation could result in the creation of fictitious states where addictive behaviours would always have high value. In gambling, due to the uncontrolled alternation of high rewards and punishments, such fictitious trumped-up states would likely appear. However, this approach still did not consider conditions characterised by information that is readily accessible to the agents, but that require substantial exploration or exceeds agents' computational capabilities, therefore resulting in incorrect representations of the environments and action sequences execution consequences.

Similar to Redish and colleagues' work on missing information [9], our study highlights a vulnerability that does not require a pharmacological interaction and derives from the cognitive and computational demands posed by the interaction with non trivial environments. This vulnerability builds on two closely related factors that have been mostly neglected in addiction computational modelling: environment complexity and exploration. Both these factors are well recognised for their importance in other fields, e.g. dealing with environment complexity is the core problem faced by Artificial Intelligence (AI) [42, 43], and exploration is central for the AI RL as well as developmental and computational neuroscience, especially in terms of the exploration-exploitation trade-off [44–56]. It is generally accepted that environment complexity and stress hinders cognitive performance and may lead to maladaptive decisions [57–63]. In fact, the assessment of action-outcome ramifications represented in a complex internal model requires significant cognitive resources and is negatively affected by any impairment of cognitive capabilities, as for instance due to anxiety or stress [58, 64–66]. Furthermore, a complex environment requires a substantial amount of exploration and training [37, 46, 55] to be correctly represented in a useful and precise model. Recent work has shown that environments rich of stimuli, novelties, intrinsic [37, 67–69] and healthy rewards may be therapeutic for addiction [70]. However, most investigations in computational models of addiction, have focused on environments characterised by the presence of easy to compute outcomes, thus neglecting an interesting element of environment complexity. As a result, this simplification has distanced the computational models from the clinical evidence, which has long considered environmental factors, and social interactions in particular, to have a strong impact on addiction development and maintenance [71–73].

Our simulations show that the development of addictive behaviours may be supported also by specific features of the environment structure due to its effects on both habitual and deliberative processes and exploration [37, 55, 74, 75]. Specifically, under conditions in which learning to obtain an addictive reward is much simpler than

evaluating the long term negative outcomes associated with it, then addictive behaviours may emerge and persist for an extended period of time even if the decision system is undamaged. This vulnerability can be seen as the other side of the same coin of dealing with missing information [9, 76–78]. In this case, information exceeding an agent cognitive capabilities would be lost to it [79], as the case of punishments presented under disparate conditions after a rewarding addictive activity. We propose this vulnerability complements and interacts with previously described ones, capturing the emergence of addiction in the absence of pharmacological factors or stochastic alternation between high punishment and reward.

Materials and Methods

Agent. The behaviour of the agent used in the simulations (Fig 1) is controlled by a hybrid (or dual) model system [21, 32]. As such it aims to optimise its behaviour, i.e. it maximises the expected cumulative reward it receives, by simultaneously learning, through a model-free (MF) component, and computing, through a model-based (MB) component, an optimal action strategy, or policy π.

MF algorithms, such as Q-Learning [80, 81] and actor critic architectures are usually adopted to model habitual processes and resulting habitual behaviours [3, 32, 82–87]. These algorithms update the estimation of the value of an action in a state $\tilde{Q}^{MF}(s, a)$ after the action is executed, and depending on the obtained rewards and observed states. Thus their flexibility may be limited but they are computationally efficient. The MB components are instead adopted to implement planning processes [14, 29, 30, 46]. These maintain an explicit *internal model* of the causal relationships experienced in the environment, including information about state transitions following the execution of an action, and the associated received reward. This internal model is then used for the *online* evaluation of action-state sequences as if planning by navigating an internal decision tree. These processes are quite demanding in terms of computational resources but allow a high degree of flexibility. The features of the two classes of component makes them complementary and several *dual* or hybrid models that integrate them have been proposed in the literature [14, 83–85, 88–91].

A key issue of most dual models is that they assume an ideal MB process [88, 92]. Such ideal MB systems have complete knowledge of the environment and unlimited computational resources, thus they can always make optimal choices when in control. These theoretical conditions are of limited interest when modelling biological phenomena and they conflict with several constraints, or bounds, that apply to any physical MB system [59, 60, 63, 93–97]. Thus, it is possible that these assumptions are hindering our ability to model biologically plausible healthy and dysfunctional (as e.g. in addiction [3, 21]) behaviours. The MB component in the proposed model drops these ideal assumptions. It learns the statistics of the environment by explicitly keeping count of the transitions and rewards observed during interaction and it is *bounded* [63, 93] in terms of the amount of computational resources it employs to estimate the best next action. At each step this component starts from scratch the value estimation (Q^{MB}). It applies the Bellman Equation a limited number of times to states sampled stochastically, following a heuristic for efficient state update selection. The algorithm is an early-interrupted variation of the Prioritized Sweeping [99], with stochastic selection of the states to update (see Algorithm 1 and for more details check the supplementary materials). Prioritized Sweeping algorithm has been chosen, like in [21], because of its sensitivity to the reward function over the whole environment, e.g. distant high rewarding states, and because its performance varies smoothly with the number of updates even if the environment presents several loops and sparse rewards. In the proposed model, the MF component of the agent has been implemented using the

standard Q-Learning algorithm in tabular form [80].

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;`
  $\forall a \quad Q(\tilde{s}, a) = \sum_{s'} p(s'|\tilde{s}, a) \left[R(\tilde{s}, a, s') + V(s')\right];$
  $M = \max_a Q(\tilde{s}, a);$
  $\Delta = |V(\tilde{s}) - M|;$
  $V(\tilde{s}) = M;$
  $\forall s \quad h(s) = \Delta \times \max_a P(\tilde{s}|s, a);$
  $H(\tilde{s}) = h(\tilde{s});$
  $\forall s \neq \tilde{s} \quad H(s) = \max\left(h(s), H(s)\right);$
end

Algorithm 1: Early Interrupted Stochastic Prioritized Sweeping pseudocode

Consistent with currently available evidence, e.g. lesion studies [14], the model proposed in this paper assumes that the MB and MF components do not share a common representations of the values. Thus, in the proposed model, as opposed to what suggested for instance in [21], there is no direct interaction between the MB and MF component during the computation of the respective state action values [1]. The hybrid model computes choice values balancing model-free (MF in the equations) and model-based (MB in the equations) components depending on a parameter, termed β. Six values (1, 0.8, 0.6, 0.4, 0.2, 0) have been used for this parameters to simulate different phenotypes, in a spectrum between purely model-based (β=1) and purely model-free (β=0) reinforcement learning. An hybrid value function Q^{MX} is computed by combining the values computed by the model-based and model-free components:

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

To allow for exploration, the action to execute was selected randomly 10% of the times. Otherwise, the action selected was the one maximizing Q^{MX}. The choice of this standard stochastic ϵ-greedy *persistent exploration* strategy [100] is motivated by the assumptions that the agent will never be able to reach the optimal strategy, as the world is too complex and dynamic [9], thus it continues to perform exploratory actions with a non-zero frequency. Other classical choices for the exploration mechanism are the use of a softmax [100] function (Boltzaman exploration) or the use of state visit counters and other form of intrinsic motivation factors [44–56]. However, the use of the softmax function has several theoretical issues [56, 100] and would also strongly depend on the use of the beta value, specially because the two learning components have different learning speed. Thus, the ϵ-greedy was adopted because of its minimality and common use.

Computational resources and the balance between MB vs MF represent some of important dimensions that determine individual learning and behaviours. In this study we model how they can impact on the risks and levels of addiction.

[1] This modelling decision cuts out the effects of internal replay by which MB can affect a MF system, for which evidence is growing. However, this interaction does not present a clear path for drug effect transfer from MB to MF if, as suggested in [14], the values are not shared and if the reward model does not simulate the pharmaceutical effects of drugs [21]. For addictive behaviours not involving drugs this path is then not crucial. On the other side it prevents propagation of drug effects from MF to MB through sharing values [21], as this is not currently supported by evidence [14].

Environment. We hypothesise that addictive behaviours, characterised by the *persistent and repeated* choice of a limited set of suboptimal actions, can be induced by an environment structure where high immediate reward follows the acquisition of an addictive reward (e.g. drug consumption), whereas limited punishment (negative effects of the addictive behaviour) is both unavoidable and distributed across many states following the acquisition of the addictive behaviour with stochastic peaks of severe negative effects, representing serious issues caused by the addictive reward (e.g. losing one's job). The abstract environment used in the experiment (Fig 2) reflects these assumptions and is designed to offer this kind of suboptimal behaviours together with optimal ones while keeping a degree of complexity that makes exploring the different alternatives non-trivial.

The environment (Fig 2) has three types of states: healthy reward, neutral and addictive (suboptimal) reward. The agent deals with environment variations meant to simulate two conditions: safe and addictive condition (Fig 2.A, details in the supplementary material). Under the addictive condition ($d_{drug} = 1000$ steps) the agent is rewarded by accessing both the healthy reward state (moderate reward, $R_g = 1$), and the addictive segment (high reward, $R_d = 10$). Conversely, under the initial safe condition ($d_{init} = 50$ steps), the agent does not receive any reward or punishment by entering the addictive reward (see the supplementary material for details). Table 1 shows the number of updates that the original Prioritized Sweeping algorithm uses to find the optimal policy in each phase, which are two orders of magnitude larger than the updates allowed to compute Q^{MB}.

Fig 1. Dual Learning Agent. The simulated subjects decision making architecture is composed of two parts. (i) a model free component updating model free action values using value prediction error computations, and (ii) a model based component containing a model of the environment and a planner that updates the model based action values. The two form of action values are combined linearly according to the β balance parameter to drive action selection.

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

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

Results

The exposure of simulated subjects to the environment depicted in figure 2 results in qualitatively different, almost opposite, behaviours even with identical parametrisations (the β MB/MF balance) (see Figs 3). On an agent-by-agent level, the behaviour trajectories show a bifurcation between:

healthy high (saturated) visit frequency to the healthy target and almost no visit to the addictive reward state;

addicted almost no visits to healthy target and high visit frequency to the addictive reward state.

This result is consistent for all values of β.

The *healthy behaviour* is a stable condition reached after less than 300 steps by all subjects showing this kind of trajectory. On the other hand, *addiction trajectories* is a

Fig 2. Descriptive representation of the environment. The states are disposed in a linear arrangement: on the left a state associated with a healthy reward, on the right a state associated with an addictive reward (e.g. drug consumptions or gambling), 14 states associated with the addictive reward aftereffects and in between 6 neutral states. The agent can freely traverse between nearby neutral states. The agent has to reach one of the two borders of the neutral states segment, to have access to either the state granting the healthy reward or the state associated with addictive reward and then enter the segment of states associated with the addictive reward. Entering the healthy reward state results in a moderate reward ($R_g = 1$), after which the agent returns to the the central neutral state (number 4). Entering the addictive reward state from the neutral states provides an immediate high reward ($R_d = 10$). The following after-effects states result in negative outcomes (-1.2) and actions are more stochastic making difficult to leave this set of states. The overall average punishments received in the after-effects area make transitioning to the addictive reward state suboptimal (see optimal policy and action values in the supplementary materials). The present graphical representation is meant to illustrate only the most relevant transitions, where the states 1,4,15 and 20 show a detailed description. Line width represents related transition probability value. Colour represents the action class (a_s, a_g, a_w, a_d), which matches with the transition probabilities described in the supplementary material. Neutral states area are navigable with actions a_{s2-7} which are deterministic for adjacent state while have high chance of failing for distant states. From the neutral state the agent can reach: (i) the healthy reward state if it executes action a_g, whilst being in state 2; and (ii) the addictive reward state (number 8) , if it executes action a_d whilst in state 7. From the healthy reward area, the agent can issue again a_g receiving a reward of 1 and going back to the centre of the neutral area, state 4. The agent enters the addictive area from state 8 and receives a reward of 10. Action results in the addictive area are less predictable and result in randomly switch to other nearby states in the addictive area. Only in state 15 the action a_w has higher chance of leaving the addictive area. Leaving the addictive area has a cost of -4 while every other transition inside the area costs -1.2. For an exact environment description see Table 2.

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

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

Fig 3. Evolution of healthy target/addictive behaviour ratio and std for addicted and healthy subjects. The data are separated by MB/MF factor β and by addicted/healthy condition. Addicted agents visit addictive reward state 8 more often than the healthy target state by definition for the whole experiment duration. The population consisted of 900 (healthy+addicted) subjects for each β value (5400 total simulated subjects). Each display reports the mean and standard deviation of the ration between the visits to the addictive reward and the sum of the visits to both the addictive and healthy reward in time interval finishing at that point and starting at the previous one. Addicted agents gradually stop visiting the healthy reward state. Conversely, healthy subjects achieve a ratio between addictive and healthy condition visits near to 0. Healthy agents with extreme values of β (0 or 1) are slower or reach lower healthy target visit frequency.

long-lasting transiently pseudo stable condition reached after less than 100 steps. In the long run, simulated addicted agents' behavioural trajectories can still converge to optimal healthy behaviour. This was shown in longer simulations where convergence happens after around 100K steps (Fig 5), or more than 300 times the number of steps required for the healthy agents and 1000 times more than the time necessary for the addictive behaviour to take place. In these simulations we employed agents controlled by MF system only. The presence of this critical bifurcation with final convergence in purely MF agents is useful to certify that the vulnerability of the hybrid agents to environment complexity was not caused by two potential computational issues. More precisely: (i) since the final convergence is a feature of the adopted undisrupted TD learning algorithms [80], the long-lasting, but transient, divergence observed for the purely MF agents in Fig 3 may have resulted from an implementation or computational error (e.g. due the limited numerical precision). If that were the case, the divergence observed in hybrid agents (Fig 3) would have been caused by such error, but no final convergence should be observed;; (ii) another possible cause of divergence in hybrid agents could have been found in the bounded MB component. This component does not have any theoretical convergence property due to its bounded resources and the limited exploration, which both result in a poor model of the world. As a consequence, the MB component may maintain its preference towards suboptimal decisions even after a long training time, thus also indirectly affecting the estimation in the MF component learning, and preventing the agent from abandoning the addictive reward related policy. Both these possible issues are proved wrong by the results represented in figure 5, where bifurcation between healthy and addicted trajectories appears while maintaining the final convergence even in absence of the bounded MB component. There are strong theoretical bases to explain this extended suboptimal behaviour of the MF as its convergence is only guaranteed after an arbitrarily long time [80]. Furthermore, from Fig 3 it appears that MB component would have helped directing search and allow a convergence in less than the hundreds thousands of steps observed in Fig 5. An

extended test was not performed due to the excessive simulation time that would have been necessary given that a step of the MB component is more than one hundred times more expensive than one of the MF component. Still in Fig 3 the addiction phase is established in about a hundred steps and lasts at least 10 times more, a non-negligible interval.

Fig 4. Ratio and confidence intervals of addicted subjects varying β, i.e. different balances between MF and MB components. Total population 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. Healthy subjects show the inverse behaviour. Intermediate values of β showed a lower ratio of addicted agents. Confidence interval were estimated assuming two-tail distribution and 95% confidence.

Fig 5. Long runs with logarithmic time scale. Results show behaviour expressed by purely MF agents ($\beta = 0$). Average on 100 runs divided between those that show behave like addicted **in the first 10,000 steps** and those that behave like healthy. Similarly to figures 3 the behaviours of the 2 groups are substantially different with net preference for addictive or other rewards. After 200K steps most of the addicted subjects go back to healthy, whereas addiction requires few tens of steps to initiate. This is shown also by the histograms in subplot B and C that show that most of the agents fall in the two opposite possible preference, in the interval with lowest drug intake ratio (0-0.125) or the interval of highest intake ratio (0.875-1). This shows that the agents are mostly stable in these two conditions. During this extremely long experiment, some agents behaviours converges to optimal and switches from the unstable high ration condition (addiction) back to the healthy stable condition.

Simulated subjects that employ MB and MF components in a balanced manner show a higher resistance to addiction in comparison with those characterised by unbalanced control (40% for $\beta = 0.4$ VS 60% for $\beta = 0$, i.e. purely MF, and 51% for $\beta = 1$, i.e. purely MB) (see Fig 4). This is likely a result of the increased robustness induced by having two parallel decision systems.

Figure 3 also shows that healthy purely MB agents achieve a lower amount of healthy reward. This is due to using a bounded agent which cannot consistently compute the optimal action selection even with a correct model. Moreover, the internal model acquired is imprecise due to limited exploration and environment complexity.

Fig 6. Ratio of addicted and confidence interval with different amount of cognitive resources (Model Based Updates per Step (MBUS)) and MB/MF balance factor β. Initial performance show a significant level of addiction for all values of β and cognitive resource, with a significant low for very high resources (500MBUS) and for a $\beta = 0.25$. On 10K steps the low resource and high resource diverge, with lack of addiction with very high resources and $\beta = 1$. Balanced conditions are more stable then extreme for most of the parametrizations. The difference between figure A and B shows the effects of the exploration on all the parametrizations. Confidence interval were estimated assuming two-tail distribution and 95% confidence, 100 simulated subjects employed for each value of β.

To converge to optimum, when the model of the environment is known, the prioritized sweeping algorithm used in the MB requires above 4K updates or iterations (Table1) of the value function using the Bellman equation (eq. **??**). Simulation results are in accordance with data showing that an impaired MB component may increase the

chances of addiction (see Fig 6). In fact, a reduced number of updates in the MB component (50 Model Based Updates per Step, MBUS) results in 50% of the subjects addicted after 1K steps rising to about 90% of the subjects after 10K steps. A very efficient MB component does not totally avoid addiction. Even with extremely high computational resources (500 MBUS) [2], 20% of the subjects show addiction at 1K steps. This percentage eventually falls to almost 0% after 10K steps, so that addicted agents do converge towards healthy behaviours, but it takes them a number of steps equivalent to 10 times the one required by addiction to take place. Furthermore, when the model was known, the selected amount of computational resources (500 MBUS) did not result in substantially worse performance (about 96%) than the optimal agent which stops at convergence (Table1). The change of performance, and the reduced addiction, results from both acquiring a correct model and consistently computing an efficient behaviour. Consider that, even if the environment is more complex than standard minimal environments used in previous models of addiction, its structure is still rather simplified in comparison with an ecological set-up. It is reasonable to assume that, in a the real world environment, the computational resources and the amount of exploration necessary to obtain similar results would be prohibitive. Thus, high computational resources and strong deliberative capacities, can help avoiding addiction, but a purely deliberative agent with high cognitive capabilities can still be susceptible of addiction due to exploration issues.

Fig 7. Effect of pre-training time (PTT, timesteps in safe phase) and MB/MF balance factor β. Exposure to the environment before the introduction of the addictive reward decreases the probability of addiction for all the subjects parametrizations. Extreme parametrizations ($\beta \in \{0, 1\}$) have still a residual tendency to addiction even with long exposure. 100 samples for each condition. Experiment length 10K steps. Evaluation on the last 1K steps.Confidence interval were estimated assuming two-tail distribution and 95% confidence, 100 simulated subjects employed for each value of β.

The only factor that can lead to the observed behavioural bifurcation, i.e. two identical simulated subjects in the same environment to show diverging behaviours, is the stochastic action selection which is dominant during the initial exploration phase of limited knowledge. In fact, at the beginning Q-values are not informative. Initial choice and the early discovery of the rewarding states shape the initial values estimation biasing future exploration. The role on initial experiences in determining extended addictive behaviours was already pointed out in [**?**] and will be further discussed later.

Early acquisition of a correct model of parts of the environment that comprises the healthy rewards and a policy to achieve them, should reduce this initial bias and improve resistance to addiction. This is confirmed by the experiment reported in Fig 7. In that simulation the subjects were allowed to explore the environment before the introduction of the addictive reward for 50, 200 and 1000 time steps (Pre-Training Time, PTT). Higher PTT, substantially decreased the chances to develop addiction independently of the balance β between MB and MF, but with balanced systems always more resistant to addiction. This result may provide the theoretical basis for current and novel prevention strategies.

The same reasoning explains why sudden environment changes are dangerous conditions for igniting addiction and relapse [64, 66]. In some preliminary simulations we tested the effect of alterations in the environment after the consolidation of a

[2] It has been noted that a possible mechanism for MB implementation is conscious mental time travel based on theta cycles in hippocampus. In this case, the updates would be very rare and probably in the order of single digits. In fact theta cycles are running 7 times a second, which could produce a few hundred in the course of a typical human decision [31, 101, 102].

Fig 8. Effects of environment change on healthy subjects After time step 2500 the healthy reward state is changed. Most of the healthy purely model based agents ($\beta = 1$) switch directly to the addictive behaviour. Agents with a non zero MF component unlearn the behaviour to reach first healthy reward location and gradually explore the addictive and the second healthy reward location. Rise of visit of the first healthy reward is due to avoiding the previously rewarded action (from healthy reward to the starting state, see Fig 2) and exploring cyclic behaviours around the first healthy reward state.

behavioural policy, thus requiring the agents to learn a new goal directed strategy. In this case the environment is similar to that shown in figure 2 but the neutral area has a Y shape. In one corridor the addictive reward is present. The other two corridors lead to possible healthy rewards. Until time step 2500, only one of the two corridors contains a healthy reward, and after reaching it the agents are sent back to the initial position. After time step 2500, the reward is switched to the other healthy reward corridor and when the old healthy reward location is reached, the agent is not sent back to the initial position but can freely move to the adjacent neutral states. The results show that after the environment variation, healthy subjects may fall into addiction (see Fig 8). This result is induced by the specific environment structure used in this scenario, but it proves that it is possible to model the shift to addiction induced by environment changes.

Discussion

A common neuropsychological view of addiction assumes this behavior is dominated by habitual decision making and inelastic behavior in the face adverse consequence (or reduced devaluation), caused by the pharmacological effects of drugs on the dopamine system [15, 103–105].

The seminal work of Redish [19] was one of the first attempts to employ RL to model addiction. It showed that a purely MF agent could be hijacked by 'adding a noncompensable drug-induced dopamine increase', corresponding to a positive prediction error. This formulation hypothesized that drug values are always underestimated by the learning system, leading to unbounded growth of the value of drug related actions. The resulting simulated behaviour reproduced several key features of addiction, but it also led to predictions whose validity is still controversial. Notably, this mechanism predicts that for drug related stimuli no *blocking* phenomena [106] should appear. However, blocking has been shown to appear both with natural and drug rewards [23, 107]. Even when a subset of high responders did not show blocking to nicotine, they were found to show blocking to food [108]. These opposite results suggests that the absence of blocking could be limited to specific phenotypes and drug rewards or that blocking may appear in early phases of addiction. Further, when the 'blocking problem' found a computational solution based on averaged rather than discounted values [23], the resulting model did not account for the devaluation of drugs when followed by aversive effects (so called 'punishments'), as observed in the early stages of addiction [22, 109, 110]. Finally, Keramati and Gutkin [22] proposed a hierarchical RL that accounts for both blocking and devaluation. Similar to Redish's model [19], drug induced positive prediction error is present, with the difference that its strength is amplified when going from high to low levels of the hierarchy [84]. By these means, the authors propose to explain the apparent conflict between cognitive and habitual evaluation of drug-related behaviours in addict individuals.However, Keramati and Gutkin's model [22], cannot generalise the overevaluation of drug-seeking behaviours in

presence of drug related cues to situations which lack such cues, where the MB would have to dominate. Low-level components are dependent on context and generalise less, whereas high-level MB ones have a broad activation context and evaluate negatively drugs consumption. So the high level would likely avoid any choice which would involve future drugs consumption and thus the activation of any low level component which leads to such behaviour. Thus, this model and the previous ones, presenting only a malfunction related to MF components, do not explain non-habitual drug-seeking behaviours and drug-seeking behaviours expressed in absence of drug related cues. They also do not account for addiction-like behaviours expressed in non-drug related conditions, such as internet gaming and gambling, which lack the consumption of substances that are assumed to hijack the dopamine-related prediction-error mechanism.

An alternative approach proposed by Simon and Daw [21] was a modified version of DYNA, one of the earliest dual models in AI [14, 90, 111, 112]. The original DYNA architecture assumes a shared value representation that would be updated by both the MB and MF process. This value sharing allows the modified DYNA model that Simon and Daw propose to overcome one of the issue in explaining the effects of drugs on MB process. In fact, value sharing allows transferring drug-related vulnerabilities of the MF dopamine system to the MB process. However, current data, e.g. robustness to lesions, indicates that the MB and MF systems, while interacting, could be more independent than it would be expected if they shared a common action value representation [14], probably cutting also this interaction path. Here we propose to address these shortcomings of computational models of addiction by investigating putative systematic failures in the MB component. We assumed a relatively complex environment, explored by an agent characterised by a bounded MB component would result in a failure in correctly estimating positive and negative outcomes associated with the available courses of action. Our simulations show that, when punishment is harder to learn than reward, an agent, with limited exploration or bounded cognitive resources, can mistakenly over-learn to chase obvious rewards that lead to non-obvious punishments. This failing can result in addiction-like behaviours even in the absence of an explicit drug-like substance.

Computational models of addiction are often characterised by reduced environment complexity and, more importantly, minimal exploration costs. Recent work on animal models has shown that environment rich of stimuli, novelties, intrinsic rewards [37, 67–69] and actual reward may be therapeutic for addiction [70]. However, addicted individuals in natural, non-controlled, environments face difficulties in acquiring healthy rewards while exploring and interacting with vast unrewarding regions. To the best of our knowledge, computational models did not investigate the effect of high exploration cost and difficult to estimate negative consequence of addictive behaviours. The interaction with such complex scenarios necessarily involves several components of the decision system, such as attention control, Pavlovian, and state identification [3, 9], with the deliberative system engaging in complex goal seeking behaviours that the habitual system alone would not be able to generate. In such environments, the adverse consequences of pursuing addictive behaviours may be unpredictable, mostly of limited intensity, spread over multiple states (e.g. tiredness) and distributed over long periods of time while presenting stochastic peaks of high intensity (e.g. losing ones job). This structure of outcomes would reduce the chances of an individual to form a comprehensive model and predictions of the consequences associated with the addictive behaviour. Our simulations are consistent with this hypothesis. We show that addiction-like behaviours emerge and persist for a long time independently of physical failures of the learning mechanisms, but rather as a result of the complexity of the environment and the associated functional shortfall in exploration and complex deliberation. This idea complements the theory proposed in [9] to explain

gambling as it focuses on a different type of suboptimal addiction-like behaviours, that are not characterised by stochastic alternation between high reward and punishment. This may be the case of internet gaming where reward is limited but regular while negative effects are delayed and less predictable, e.g. social withdrawal [10].

Contrarily to the vulnerability described in [9], which derives from errors in identifying the current state or situation, the vulnerability proposed in our model may be seen as an error in identifying or estimating the environment as a whole. In RL literature, this corresponds to a vulnerability in the active process of exploration [44–56]. Our simulations show that, depending on less than *few hundreds* initial (stochastic) exploration steps, the simulated behaviour of the agents bifurcates. Eventually, two possible behaviours emerge, as the agents consistently and repeatedly either select the healthy rewards in the environment or perform a maladaptive selection of the suboptimal addictive rewards outcomes. Importantly, the latter behaviour may last up to *hundreds of thousands* of steps. The role of chance in this process is similar to the one suggested for gambling in [9]: 'A sequence of unexpected wins can produce anomalous expectations that will not be unlearned by subsequent losses..' , thus leading to the suboptimal behaviours typical of gambling. Differently from other works in the field [19, 22], the bifurcation is not due to a disruption of the classical TD-MF learning mechanism. The MF algorithm employed in the simulated agents is a standard and converging one.

Finally, we tested whether this vulnerability could be found consistently across different configurations of parameters, and we explored different values for the balance of control between MB and MF (β) to find that the bifurcation is present with non-zero probability for all tested values of β.

Crucially, this consistency across parameters confirms addiction-like behaviours in our simulations are the result of specific features characterising the interaction between the agent and the environment. In particular, such determinant features are: (i) an easy to access state with a high reward, which is is not limited to the simulation of the consumption of substances of abuse, followed by (ii) difficult to estimate adverse consequences, such as delayed, stochastic or weak and prolonged negative rewards. Any action choice leading to high immediate reward and long-term spread punishment, as in the case of addictive behaviours, can be similarly modelled.

The structure of the environment concurs in defining exploration costs, e.g. the time required to find the optimal policy. Both MF and MB decision processes show significant difficulties when the environment presents the above described features, even when the MB component has access to substantial computational resources. As shown by the simulations, these features may lead both MB and MF components to overestimate for long time the value of the actions directed to the addictive reward . In fact, the agent, while being 'lured' by the high initial reward of drug, may neglect the exploration of competing and more convenient strategies and persist in pursuing an apparent 'local optimum'. In this scenario, on one hand, the MF component has substantial difficulties in propagating the delayed punishment that follows the suboptimal goal, as diverse states are traversed after reaching the suboptimal target while negative reward is stochastically assigned to the agent. On the other hand, the MB component is resource bounded and may fail in evaluating long term action effects under such conditions in which punishment is limited with peaks of intensity spread across multiple states. Thus, even if the MF component can eventually converge after extensive exploration, the MB component may keep pursuing sub-optimal addictive behaviours even after such extensive exploration.

We propose that the relatively complex structure of the chosen environment abstractly reflects addicted individuals' ecological social and environmental interactions. Infinite different configurations and structures can be conceived that lead to suboptimal

policies, yet, not all environments or suboptimal agents would express addiction-like behaviours. For instance, random environments tuned to assume different values of complexity, defined according to one of the several available metrics (as in [51, 52]), or dual RL agents that show a spectrum of sub-optimal behaviours during learning would not likely lead to the repetition of suboptimal selections followed by negative outcomes. Our approach appears more fit to model addiction. Our results highlights that the use of oversimplified environments in previous computational studies may have led to neglect the limits of the MB components and to overemphasize the role in addiction of an overactive MF component. Under such conditions, a MB component is unlikely to cause suboptimal behaviours as values are easy to estimate and update. Conversely, a richer than usual environment facilitates the analysis of the possible dysfunctions of MB decision processes and associated disorders in planning or exploring.

Our model shows that, even in the absence of any pharmacological malfunction induced by drug consumption the following limits affect a MB system:

1. MB relies on costly computations and has computational limits [59, 60, 95, 96]. The evaluation of long term consequences (value) of addictive behaviours may not be trivial. The same computational limits may result in an extreme under-evaluation of non-addictive-behaviour related conditions, making irrelevant negative long term effects of addictive behaviours. Importantly, this source of inefficiency, differently from those of the MF process, would *not* be ameliorated by exposure to further experience and associated learning, as it depends on the intrinsic limits of the subject's cognitive resources which might be temporarily reduced for instance due to stress or cognitive load [58, 64–66]. This availability is represented in our model by the number of updates that the MB component performs per state. An agent under transitory stress conditions would be simulated by temporarily decreasing this same number regulating the updates.

2. MB performance crucially depends on the quality of the available model of the environment. Acquiring an effective model of the environment by direct experience requires devising efficient exploration strategies, especially in terms of the exploration-exploitation trade-off [37, 44–56]. Two main objectives are proposed for this trade-off: achieving an efficient policy with a minimum number of samples (PAC [47, 49, 52]) versus minimising the overall cost of learning efficient policies (regret [51, 54, 56]). The mechanisms that regulate the exploration/exploration trade-off in humans are not well understood, but the heuristic methods are the most likely to be biologically plausible [44, 45, 75]. These issues leave MF learning and MB model acquisition, and consequently MB computations and resulting behaviours, vulnerable to (arbitrarily) sub-optimal decisions during exploration even if high computational resources are available.

A psychological theory assumes that the brain represents future events differently depending on their temporal distance, these time-dependent representations are termed Temporal Construals. As suggested in [113], they may lead to several maladaptive behaviours, such addictive behaviours, due to the imprecise estimation of future states and outcomes they inevitably will induce. Temporal construals can be seen as a strategy adopted by the brain to partially address the first limit described above, in dealing with a complex environment with bounded computational resources. For instance, future similar states may be compressed and represented in a more abstract way, thus requiring smaller computational costs for MB estimations of action values. Still, as noted above, this strategy may lead maladaptive behaviours when it fails. Indeed temporal construals will be affected by the second of the limits hindering MB, which is related to exploration and model precision. Temporal construals may actually strongly affect the precision of the model, e.g. by erroneously abstracting some states and

making inaccessible important ramifications of future scenarios to the MB estimation process. Thus, such bounded temporal representations can be an additional issue to consider jointly with the two limits described above, however for the scope of this paper, we focus on the effects of bounded rationality independent of the time component.

Other models [83, 89] simulate a contextual modulation of the balance β. Conversely, our results are consistent across a spectrum of values of β including the extreme 0 and 1, thus the observed phenomena could in principle take place for any mechanism that modulates β. This investigation into the agent-environment interaction features that favour addiction-like behaviours, can also shed a light on those interactions that are most likely to match biological data.

Our results may help explaining some inconsistencies between theory and experimental data on MB-MF balance, whether pertaining behaviour or neural activity. Some data show that a reduced contribution of the MB component may be a risk factor for addiction [114]. In our model terms, such data seem to indicate that the subjects characterised by smaller values of β are expected to be associated with higher risk of developing addiction. This is also true in our results for values of β smaller than 0.5 (Fig 4). However, multiple studies have also highlighted a significant role in addiction is also played by brain areas commonly associated with MB decision making [115–119], suggesting a more complex understanding of MB-MF interactions in addiction that matches our simulated results where the behavioural bifurcation between healthy and addicted behaviours, is found for all values of β, independent of any pharmacological disruption of the MF learning mechanism.

We hypothesise that higher robustness that our simulations ascribe to balanced MB and MF agents is an effect of the different vulnerabilities characterising the two components and their complementary strengths when facing complex environments [3]. In fact, the simulations suggest that a balanced MB and MF system may not only be more efficient in terms of computational and experience costs [90] but it may also be more resilient against different maladaptive behaviours and malfunctions (as, for instance in the case of lesions, see [84]).

Our results can be considered complementary to computational and experimental work related to the effect of the RL discount factor in addiction [113, 120–122]. Indeed, having a lower γ discount factor for both the MB and MF components, and thus giving less importance to negative consequences of addictive behaviours, could favour the execution of addictive behaviours if they result as optimal for such parametrizations. However, such conclusion could not explain why addicted agents would express the desire to quit but cannot [22], as addiction would be a properly optimal behaviour for both the habitual and deliberative systems. The same would happen if the MB discount factor were lower than that of the MF one. However, a decreased discount factor only affecting the MF component would give rise to the deliberative system trying to resist the habitual one, possibly expressed as frustration about the inability to quit, but it would not explain complex behaviours relying on the deliberative process aimed at final drug consumption or addictive behaviours (e.g. planning a trip to Las Vegas in order to gamble). Interestingly, models like those proposed in [113, 121] are also related to the bounded cognitive resources. Rather than considering the deliberation cost, in this case the complexity of beliefs is considered bounded. In [121] the authors propose that, instead of a complex belief state comprising also the uncertainty on the correct discount factor, the model employs multiple simpler MF agents each with a simpler representation. This explains several important observations such us hyperbolic discount. However, such a MF process cannot account for complex (non-habitual) behaviours aimed at final drug consumption or addictive behaviours. In general, dysfunctions due to the bounded complexity of representations would account for consistent failures of the decision process that can explain several characteristics of

addiction. Still, it may be difficult to explain the observed variation of relapse occurrence between and within subjects only on the bases of bounded representations. In our model, the propensity to randomly fail complex MB estimations, in putative relation with stress levels [66], could offer a way to explain these observations.

Our model and results have interesting implications for treatment development, but it is important to remember that the other vulnerabilities [3, 9, 19, 22] must be taken into account as well. The MB and MF components have different vulnerabilities and learning dynamics. Different forms of therapy may be more effective on each of the components. Furthermore, depending on the β balance factor they have a different impact on the development of the addictive behaviour, thus to be effective therapies must take into account such balance. The MF vulnerabilities are mostly caused by slow adaptation, or more precisely, by the difficulty of generalising new results. For a MF component, a radical change of contest with simultaneous behavioural retraining based on incentives to eradicate addictive reward seeking habits would be the most effective strategy. Conversely, the MB can rely on wider set of knowledge in comparison with the MF component, so that different forms of therapy can be devised. However, this higher level of flexibility makes the MB able to adapt and reinstate the same addictive behaviour even after radical changes. A crucial problem is that the MB cannot increase its computational power with training, so that even if a correct model is formed, the agent might still pursue addictive behaviours, initiating relapse. While a solution could be to simplify or make more explicit and accessible the association between normally occurring negative outcomes and the addictive behaviour, there is the possibility that whatever therapy is employed, MB can 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. An important step is thus to study therapies that develop addiction-avoiding habits. Such habits can counteract the failures of the MB system, specially when this is more prone to failure (e.g. due to stress) buying time to wait until conditions for MB computation become again more favourable. This strategy is supported by the results of the pre-training experiment shown in Fig 7. We propose a preliminary differentiation between balanced and unbalanced phenotypes maybe assessed behaviourally, with tasks presenting deterministic rewards and stochastic delayed negative outcomes (e.g. the Iowa Gambling Task [123]). However, an assessment of MF or MB dominance may require a more sophisticated analysis of the effective connectivity between front-ventral regions (often associated with MB) and dorso-lateral regions (often associated with MF).

Conclusions

In conclusion, a few statistical studies focus on the effects that different sources of complexity (most prominently, social factors [124, 125] and stress [66, 126]) may have on addiction, however current computational modelling literature mostly disregards these aspects [71–73]. In this work we have proposed a step forward in the simulation of complex environments with features that putatively resemble the actual ecology of addiction.

The significant way in which different features of the environment affect the subjects' behaviours, calls for methods to extract more detailed characterization of the environment in order to better replicate any healthy and dysfunctional behaviour. Our simulations show (in particular Fig 5) a combination of two ecologically plausible conditions, namely environment complexity and bounded MB component, result in addiction-like behaviours even in the absence of learning dysfunctions. We propose that, by considering complex environment features and the associated generation of limited internal models of the world, our model reveals important missing factors in the

understanding of mental disorders in general and addiction in particular.

References

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

2. Tschernegg M, Crone JS, Eigenberger T, Schwartenbeck P, Fauth-Bhler M, Lemnager T, et al. Abnormalities of functional brain networks in pathological gambling: a graph-theoretical approach. Frontiers in human neuroscience. 2013;7:625. doi:10.3389/fnhum.2013.00625.

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

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

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

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

7. Everitt BJ, Robbins TW. Drug addiction: updating actions to habits to compulsions ten years on. Annual Review of Psychology. 2016;67:23–50.

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

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

10. Petry NM, Rehbein F, Ko CH, O'Brien CP. Internet Gaming Disorder in the DSM-5. Current psychiatry reports. 2015;17:72. doi:10.1007/s11920-015-0610-0.

11. Piquet-Pessôa M, Ferreira GM, Melca IA, Fontenelle LF. DSM-5 and the decision not to include sex, shopping or stealing as addictions. Current Addiction Reports. 2014;1(3):172–176.

12. Grant JE, Potenza MN, Weinstein A, Gorelick DA. Introduction to behavioral addictions. The American journal of drug and alcohol abuse. 2010;36(5):233–241.

13. Olsen CM. Natural rewards, neuroplasticity, and non-drug addictions. Neuropharmacology. 2011;(61):1109–1122. doi:doi:10.1016/j.neuropharm.2011.03.010.

14. Daw ND, Dayan P. The algorithmic anatomy of model-based evaluation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1655):20130478.

15. ODoherty JP, Lee SW, McNamee D. The structure of reinforcement-learning mechanisms in the human brain. Current Opinion in Behavioral Sciences. 2015;1:94–100.

16. Norman DA, Shallice T. Attention to action: Willed and automatic control of behavior. Cognitive neuroscience: A reader. 2000; p. 376–390.

17. Nadel L. Multiple memory systems: What and why, an update. Memory systems. 1994;1994:39–63.

18. Cohen NJ, Squire LR. Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. Science. 1980;210(4466):207–210.

19. Redish AD. Addiction as a computational process gone awry. Science. 2004;306(5703):1944–1947.

20. Di Chiara G. Drug addiction as dopamine-dependent associative learning disorder. European journal of pharmacology. 1999;375:13–30.

21. Simon DA, Daw ND. Dual-system learning models and drugs of abuse. In: Computational Neuroscience of Drug Addiction. Springer; 2012. p. 145–161.

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

23. Dezfouli A, Piray P, Keramati MM, Ekhtiari H, Lucas C, Mokri A. A neurocomputational model for cocaine addiction. Neural computation. 2009;21(10):2869–2893.

24. Bickel WK, Marsch LA. Toward a behavioral economic understanding of drug dependence: delay discounting processes. Addiction. 2001;96(1):73–86.

25. National Insitute on Drug Abuse. The Science of Drug Abuse and Addiction: The Basics; 2016. Available from: `https://www.drugabuse.gov/publications/media-guide/science-drug-abuse-addiction-basics`.

26. Tiffany ST. A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. Psychological review. 1990;97(2):147.

27. Root DH, Fabbricatore AT, Barker DJ, Ma S, Pawlak AP, West MO. Evidence for habitual and goal-directed behavior following devaluation of cocaine: a multifaceted interpretation of relapse. PloS one. 2009;4:e7170. doi:10.1371/journal.pone.0007170.

28. Gu X, Filbey F. A Bayesian Observer Model of Drug Craving. JAMA psychiatry. 2017;.

29. Redish AD, Johnson A. A computational model of craving and obsession. Annals of the New York Academy of Sciences. 2007;1104:324–339. doi:10.1196/annals.1390.014.

30. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. Model-based choices involve prospective neural activity. Nature neuroscience. 2015;18(5):767–772.

31. Kaplan R, King J, Koster R, Penny WD, Burgess N, Friston KJ. The Neural Representation of Prospective Choice during Spatial Planning and Decisions. PLOS Biology. 2017;15(1):e1002588.

32. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011;69(6):1204–1215.

33. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010;66(4):585–595.

34. Schultz W, Dayan P, Montague RP. A neural substrate of prediction and reward. Science. 1997;275:1593–1599.

35. Faure A, Haberland U, Cond F, El Massioui N. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2005;25:2771–2780. doi:10.1523/JNEUROSCI.3894-04.2005.

36. Sadacca BF, Jones JL, Schoenbaum G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. eLife. 2016;5:e13665.

37. Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G. Active inference and epistemic value. Cognitive neuroscience. 2015;6:1–28.

38. Ognibene D, Baldassare G. Ecological Active Vision: Four Bioinspired Principles to Integrate Bottom–Up and Adaptive Top–Down Attention Tested With a Simple Camera-Arm Robot. Autonomous Mental Development, IEEE Transactions on. 2015;7(1):3–25.

39. Cassandra AR. Exact and Approximate Algorithms for Partially Observable Markov Decision Processes. Brown University; 1998.

40. Ballard DH, Hayhoe MM, Pook PK, Rao RP. Deictic codes for the embodiment of cognition. Behav Brain Sci. 1997;20(4):723–42; discussion 743–67.

41. Whitehead SD, Lin LJ. Reinforcement learning of non-Markov decision processes. Artificial Intelligence. 1995;73(1-2):271–306.

42. Russell SJ, Norvig P. Artificial Intelligence A Modern Approach 3rd Ed. Prentice Hall; 2010.

43. Geffner H, Bonet B. A Concise Introduction to Models and Methods for Automated Planning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2013;8(1):1–141.

44. Schmidhuber J. A possibility for implementing curiosity and boredom in model-building neural controllers. In: Meyer JA, Wilson SW, editors. Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats,. MIT Press/Bradford Book; 1991. p. 222–227.

45. Oudeyer PY, Kaplan F, Hafner VV. Intrinsic motivation systems for autonomous mental development. IEEE transactions on evolutionary computation. 2007;11(2):265–286.

46. Kearns M, Singh S. Finite-sample convergence rates for Q-learning and indirect algorithms. Advances in neural information processing systems. 1999; p. 996–1002.

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

48. Duff MO. Optimal Learning: Computational Procedures For Bayes-Adaptive Markov Decision Processes. University of Massachusetts Amherst; 2002.

49. Strehl AL, Li L, Littman ML. Reinforcement learning in finite MDPs: PAC analysis. The Journal of Machine Learning Research. 2009;10:2413–2444.

50. Veness J, et al. Reinforcement Learning via AIXI Approximation. In: AAAI; 2010.

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

52. Jiang N, Krishnamurthy A, Agarwal A, Langford J, Schapire RE. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable;.

53. Bellemare MG, Srinivasan S, Ostrovski G, Schaul T, Saxton D, Munos R. Unifying Count-Based Exploration and Intrinsic Motivation. arXiv preprint arXiv:160601868. 2016;.

54. Ghavamzadeh M, Petrik M, Chow Y. Safe Policy Improvement by Minimizing Robust Baseline Regret. In: Advances in Neural Information Processing Systems; 2016. p. 2298–2306.

55. Hills TT, Todd PM, Lazer D, Redish AD, Couzin ID, Group CSR, et al. Exploration versus exploitation in space, mind, and society. Trends in cognitive sciences. 2015;19(1):46–54.

56. Cesa-Bianchi N, Gentile C, G L, Neu G. Boltzmann Exploration Done Right. ArXiv Computer Science e-prints. 2017;.

57. Simon HA. Rational choice and the structure of the environment. Psychological review. 1956;63(2):129–138.

58. Starcke K, Brand M. Decision making under stress: a selective review. Neuroscience and biobehavioral reviews. 2012;36:1228–1248. doi:10.1016/j.neubiorev.2012.02.003.

59. Griffiths TL, Lieder F, Goodman ND. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. Topics in cognitive science. 2015;7(2):217–229.

60. Gershman SJ, Horvitz EJ, Tenenbaum JB. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science. 2015;349(6245):273–278.

61. Zilberstein S. Metareasoning and Bounded Rationality. In: proceedings of the workshop on metareasoning of aaai 2008; 2008.

62. Russell SJ. Rationality and intelligence. Artificial Intelligence. 1997;94(1-2):57 – 77. doi:DOI: 10.1016/S0004-3702(97)00026-X.

63. Russell S, Wefald E. Decision Theoretic Control of Reasoning: General Theory and an Algorithm to Game Playing; 1988.

64. Sinha R. The role of stress in addiction relapse. Current psychiatry reports. 2007;9:388–395.

65. Pabst S, Brand M, Wolf OT. Stress and decision making: a few minutes make all the difference. Behavioural brain research. 2013;250:39–45. doi:10.1016/j.bbr.2013.04.046.

66. Mantsch JR, Baker DA, Funk D, Lê AD, Shaham Y. Stress-induced reinstatement of drug seeking: 20 years of progress. Neuropsychopharmacology. 2016;41(1):335–356.

67. Singh S, Barto AG, Chentanez N. Intrinsically motivated reinforcement learning. Advances in neural information processing systems. 2005;17:1281–1288.

68. Gottlieb J, Hayhoe M, Hikosaka O, Rangel A. Attention, reward, and information seeking. The Journal of Neuroscience. 2014;34(46):15497–15504.

69. Bardo MT, Donohew RL, Harrington NG. Psychobiology of novelty seeking and drug seeking behavior. Behavioural Brain Research. 1996;77(1-2):23–43. doi:10.1016/0166-4328(95)00203-0.

70. Solinas M, Thiriet N, Chauvet C, Jaber M. Prevention and treatment of drug addiction by environmental enrichment. Progress in Neurobiology. 2010;(92):575–592. doi:10.1016/j.pneurobio.2010.08.002.

71. Heilig M, Epstein DH, Nader MA, Shaham Y. Time to connect: bringing social context into addiction neuroscience. Nature reviews Neuroscience. 2016;17:592–599. doi:10.1038/nrn.2016.67.

72. Hart CL. Viewing addiction as a brain disease promotes social injustice. Nature Human Behaviour. 2017;1:0055.

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

74. Dayan P. Exploration from generalization mediated by multiple controllers. In: Intrinsically motivated learning in natural and artificial systems. Springer; 2013. p. 73–91.

75. Gottlieb J, Oudeyer PY, Lopes M, Baranes A. Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends in cognitive sciences. 2013;17(11):585–593.

76. McCallum AK. Efficient exploration in reinforcement learning with hidden state; 1997.

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

78. Ognibene D, Volpi NC, Pezzulo G, Baldassare G. Learning epistemic actions in model-free memory-free reinforcement learning: experiments with a neuro-robotic model. In: Biomimetic and Biohybrid Systems. Springer; 2013. p. 191–203. Available from: `http://link.springer.com/chapter/10.1007/978-3-642-39802-5_17`.

79. Tishby N, Polani D. Information theory of decisions and actions. In: Perception-Action Cycle. Springer; 2011. p. 601–636. Available from: `http://link.springer.com/chapter/10.1007/978-1-4419-1452-1_19`.

80. Watkins CJ, Dayan P. Q-learning. Machine learning. 1992;8(3-4):279–292.

81. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Sutton RS, Barto AG, editors. Cambridge, MA: MIT PRESS; 1998.

82. Doya K. Reinforcement Learning in Continuous Time and Space. Neural Computation. 2000;12(1):219–245.

83. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience. 2005;8(12):1704–1711.

84. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron. 2013;80(2):312–325.

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

86. Dezfouli A, Balleine BW. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput Biol. 2013;9(12):e1003364.

87. Dezfouli A, Balleine BW. Habits, action sequences and reinforcement learning. European Journal of Neuroscience. 2012;35(7):1036–1051.

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

89. Pezzulo G, Rigoli F, Chersi F. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. Frontiers in psychology. 2013;4(92):1–15.

90. Sutton RS. Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In: proceedings of international conference of machine learning 1990; 1990. p. 216–224.

91. Lee SW, Shimojo S, ODoherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 2014;81(3):687–699.

92. Pezzulo G, van der Meer MA, Lansink CS, Pennartz C. Internally generated sequences in learning and executing goal-directed behavior. Trends in cognitive sciences. 2014;.

93. Simon HA. Theories of bounded rationality. Decision and organization. 1972;1(1):161–176.

94. Ognibene D, 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.

95. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012;8(3):e1002410. doi:10.1371/journal.pcbi.1002410.

96. Friston K, Schwartenbeck P, FitzGerald T, Moutoussis M, Behrens T, Dolan RJ. The anatomy of choice: dopamine and decision-making. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2014;369. doi:10.1098/rstb.2013.0481.

97. Ortega PA, Braun DA. Thermodynamics as a theory of decision-making with information-processing costs. Proc R Soc A. 2013;.

98. Huys QJ. Computational psychiatry. Encyclopedia of Computational Neuroscience. 2015; p. 775–783.

99. Moore AW, Atkeson CG. Prioritized sweeping: Reinforcement learning with less data and less time. Machine Learning. 1993;13(1):103–130.

100. Singh S, Jaakkola T, Littman ML, Szepesvri C. Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms. Machine Learning. 2000;38(3):287–308.

101. Wikenheiser AM, Redish AD. Hippocampal theta sequences reflect current goals. Nature neuroscience. 2015;18(2):289–294.

102. Kurth-Nelson Z, Economides M, Dolan RJ, Dayan P. Fast sequences of non-spatial state representations in humans. Neuron. 2016;91(1):194–204.

103. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005;47:129–141. doi:10.1016/j.neuron.2005.05.020.

104. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003;38(2):329–337.

105. McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003;38:339–346.

106. Kruschke JK. Toward a unified model of attention in associative learning. J Math Psychol. 2001;45(6):812–863. doi:http://dx.doi.org/10.1006/jmps.2000.1354.

107. Panlilio LV, 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. 2007;.

108. Jaffe A, Pham J, Tarash I, Getty SS, Fanselow MS, Jentsch JD. The Absence of Blocking Innicotine High-Responders as a Possible Factor in the Development of Nicotine Dependence? The Open Addiction Journal. 2014;7(1).

109. Vanderschuren LJ, Everitt BJ. Drug seeking becomes compulsive after prolonged cocaine self-administration. Science. 2004;305(5686):1017–1019.

110. Deroche-Gamonet V, Belin D, Piazza PV. Evidence for addiction-like behavior in the rat. Science. 2004;305(5686):1014–1017.

111. Silver D, Sutton RS, Müller M. Sample-based learning and search with permanent and transient memories. In: Proc. of 25th Int. Conf. Mach. Learn. (ICML). ACM; 2008. p. 968–975.

112. Silver D, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;.

113. Trope Y, Liberman N. Temporal construal. Psychological review. 2003;110(3):403.

114. Reiter AM, Deserno L, Wilbertz T, Heinze HJ, Schlagenhauf F. Risk factors for addiction and their association with model-based behavioral control. Frontiers in behavioral neuroscience. 2016;10.

115. Nestor L, Hester R, Garavan H. Increased ventral striatal BOLD activity during non-drug reward anticipation in cannabis users. Neuroimage. 2010;49(1):1133–1143.

116. Goldstein RZ, Volkow ND. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. Nature Reviews Neuroscience. 2011;12(11):652–669.

117. Balodis IM, Potenza MN. Anticipatory reward processing in addicted populations: a focus on the monetary incentive delay task. Biological psychiatry. 2015;77(5):434–444.

118. Hommer DW, Bjork JM, Gilman JM. Imaging brain response to reward in addictive disorders. Annals of the New York Academy of Sciences. 2011;1216(1):50–61.

119. Limbrick-Oldfield EH, van Holst RJ, Clark L. Fronto-striatal dysregulation in drug addiction and pathological gambling: consistent inconsistencies? NeuroImage: Clinical. 2013;2:385–393.

120. Story GW, Moutoussis M, Dolan RJ. A Computational Analysis of Aberrant Delay Discounting in Psychiatric Disorders. Frontiers in psychology. 2015;6.

121. Kurth-Nelson Z, Redish AD. Temporal-difference reinforcement learning with distributed representations. PLoS One. 2009;4(10):e7362.

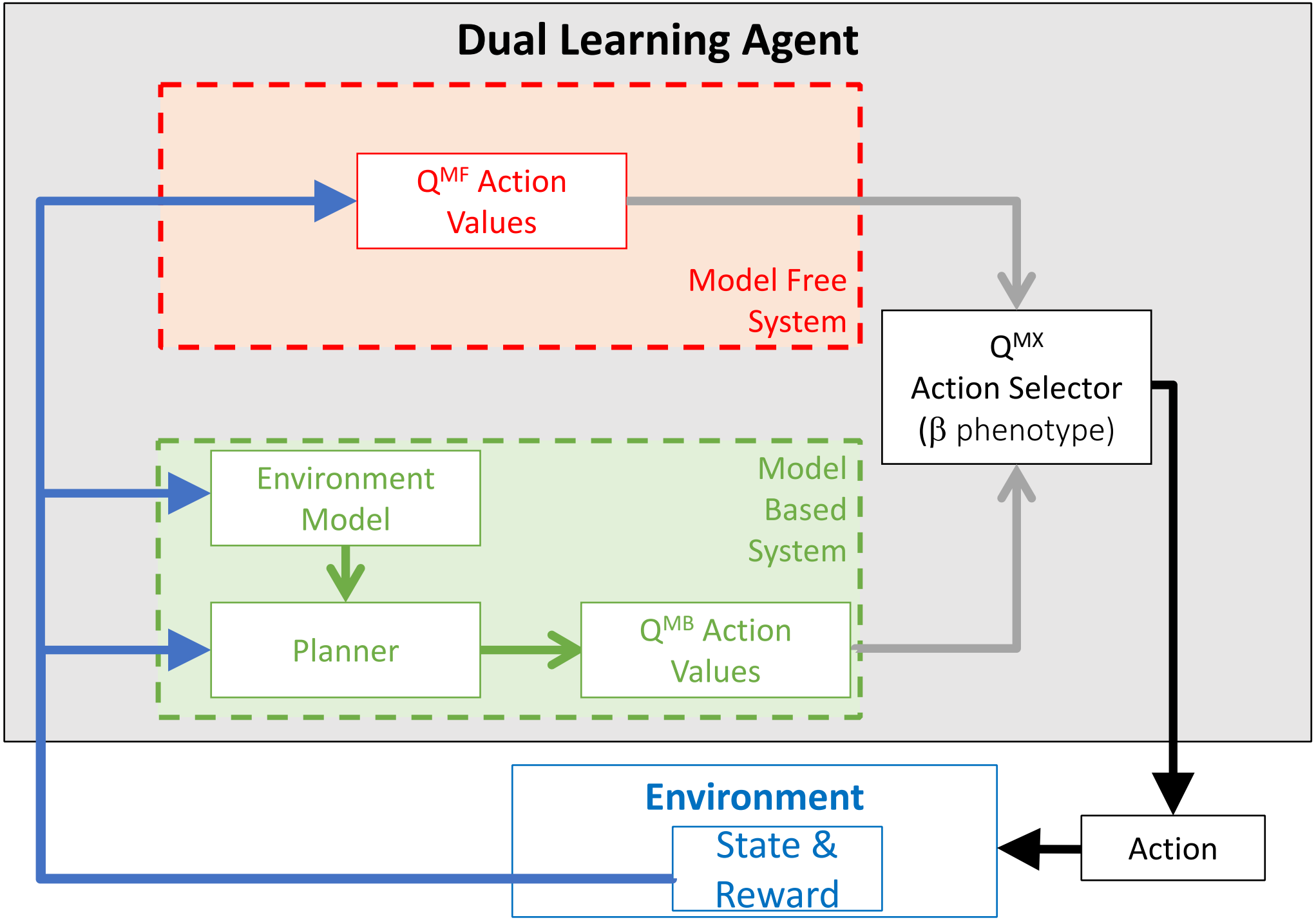
122. Kirby KN, Petry NM, Bickel WK. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. Journal of Experimental psychology: general. 1999;128(1):78.

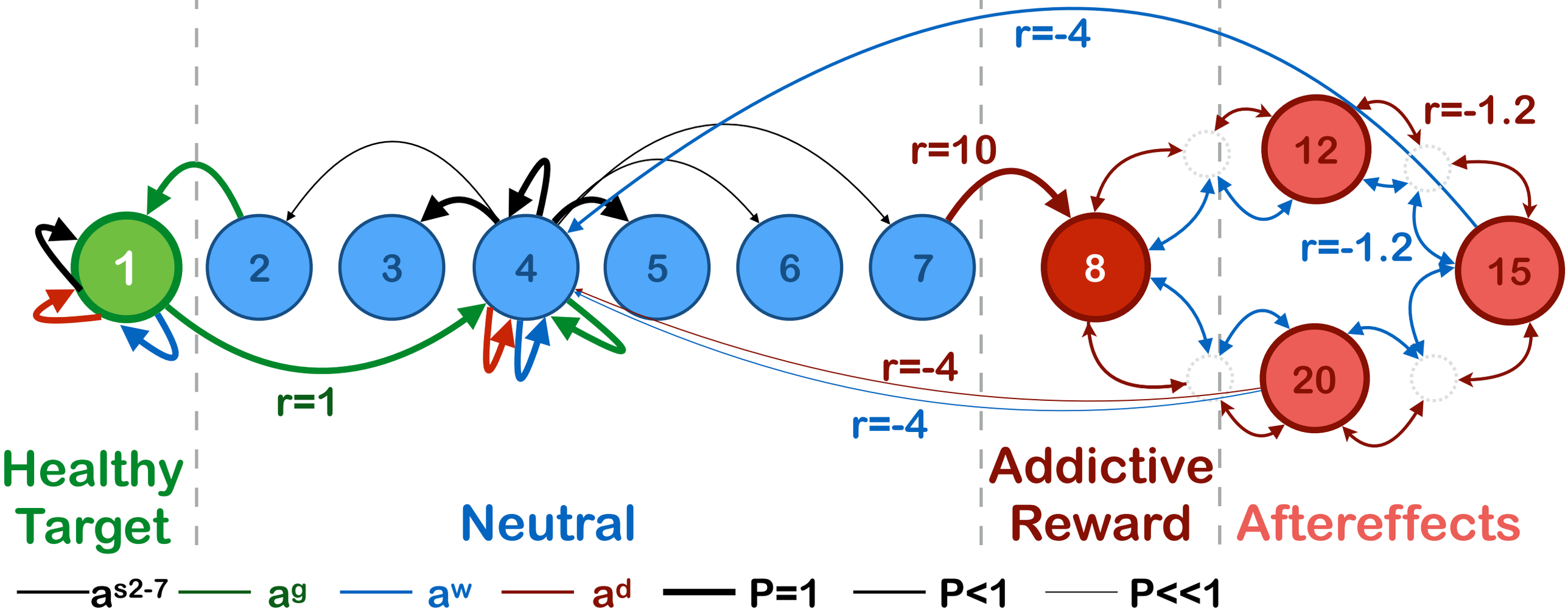
123. Dow Schull N. Addiction by design. Princeton, NJ. 2012;.

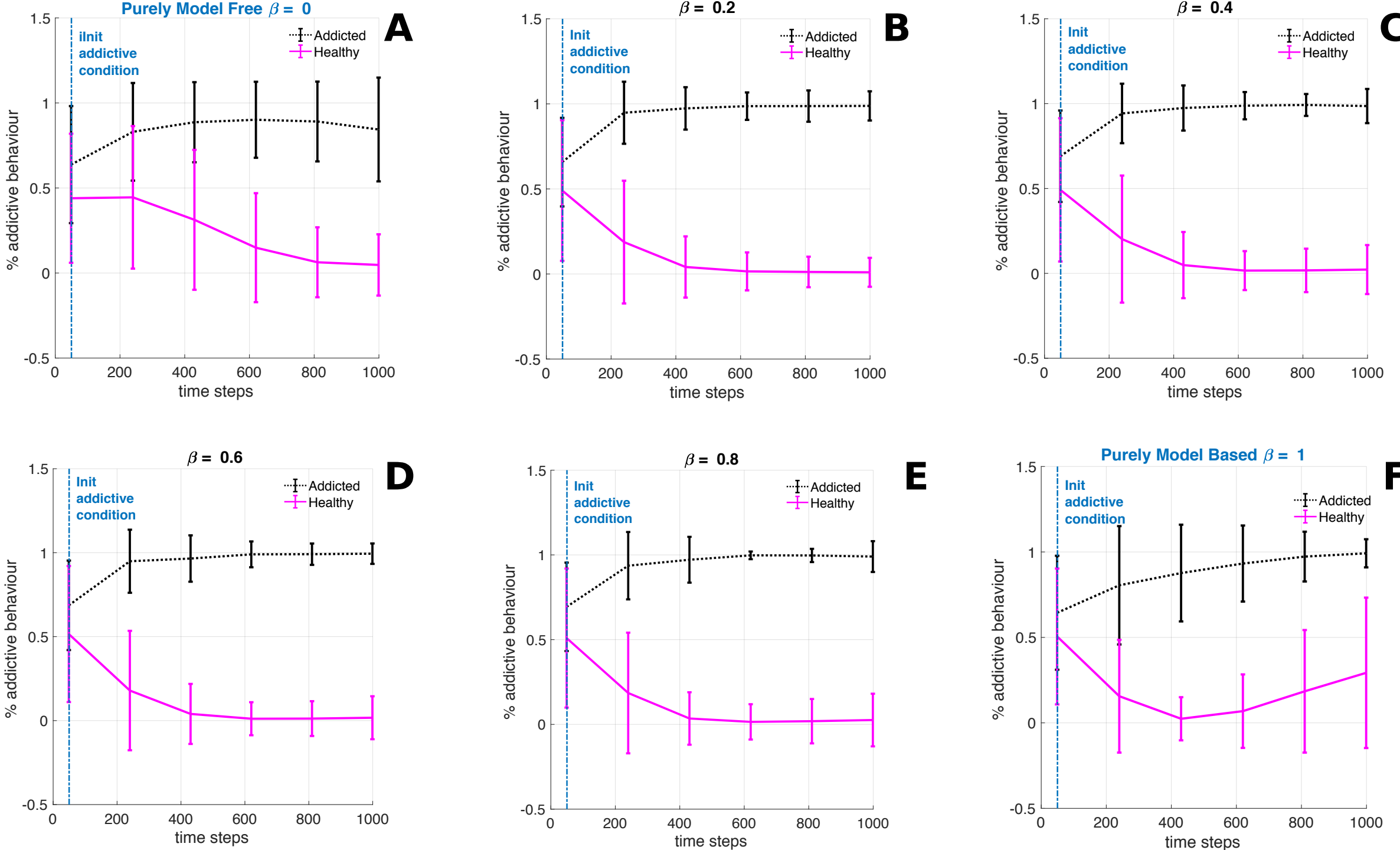
124. Berkman LF, Kawachi I, Glymour MM. Social epidemiology. Oxford University Press; 2014.

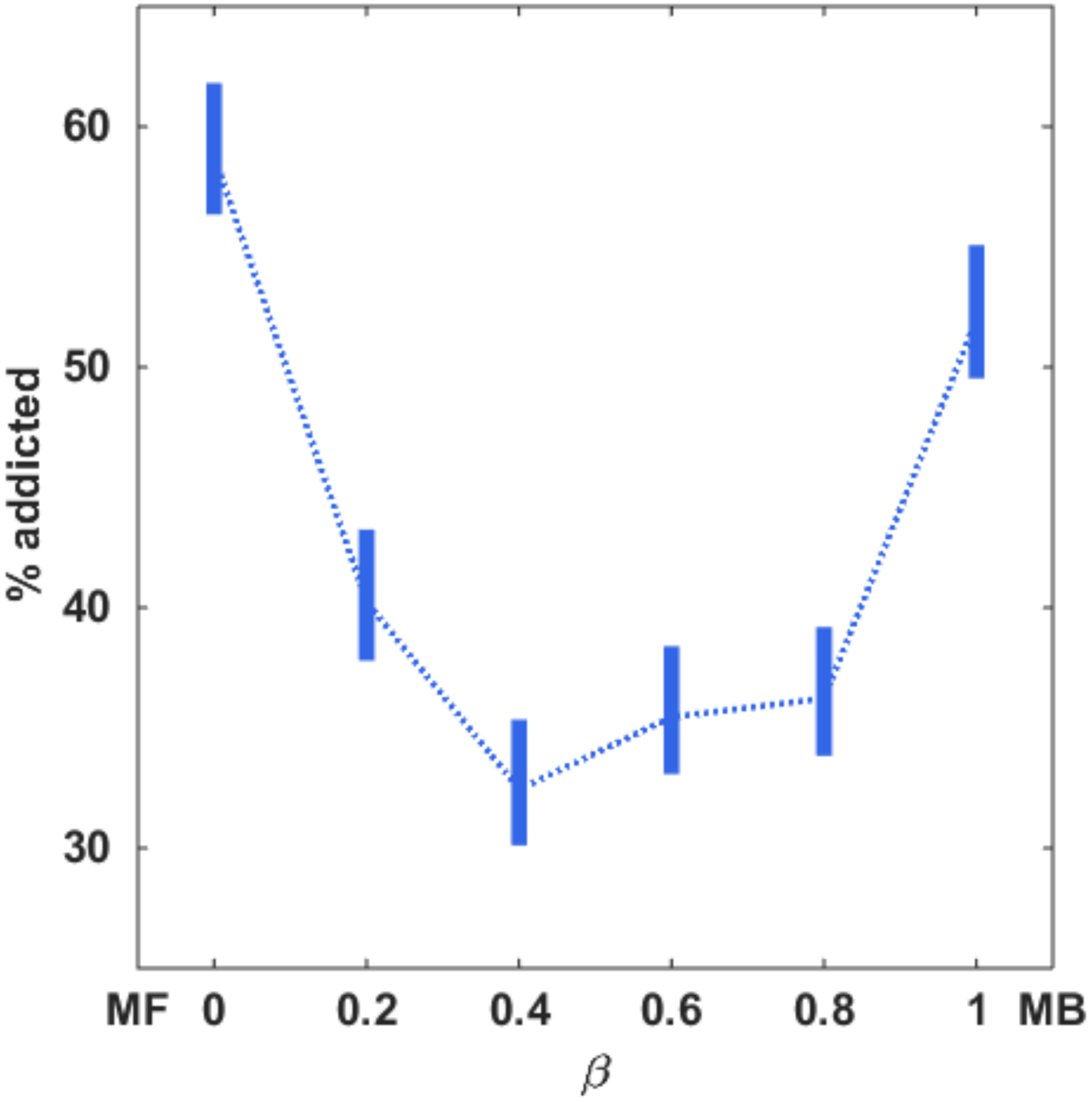
125. Havassy BE, Hall SM, Wasserman DA. Social support and relapse: Commonalities among alcoholics, opiate users, and cigarette smokers. Addictive behaviors. 1991;16(5):235–246.

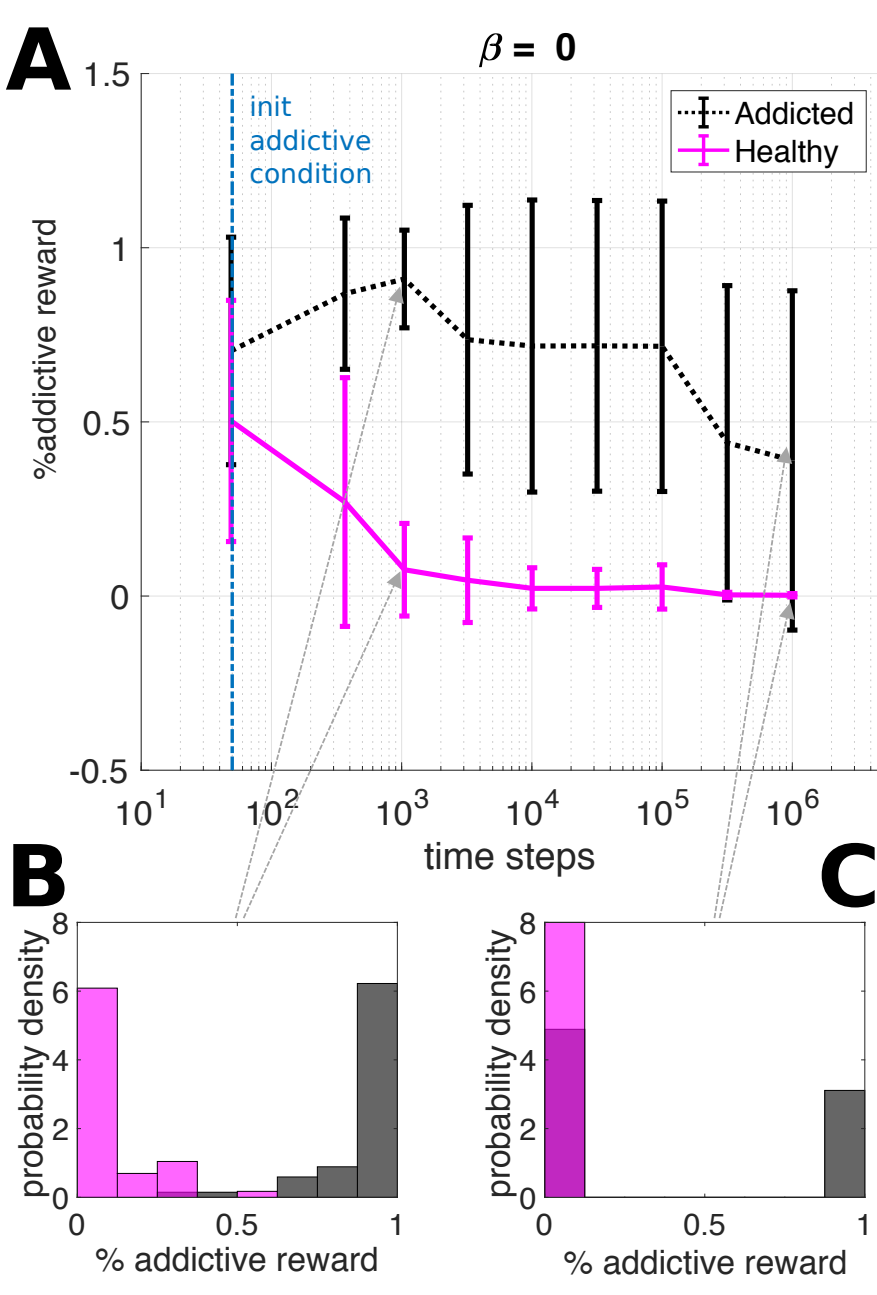
126. Sinha R, Shaham Y, Heilig M. Translational and reverse translational research on the role of stress in drug craving and relapse. Psychopharmacology. 2011;218(1):69.

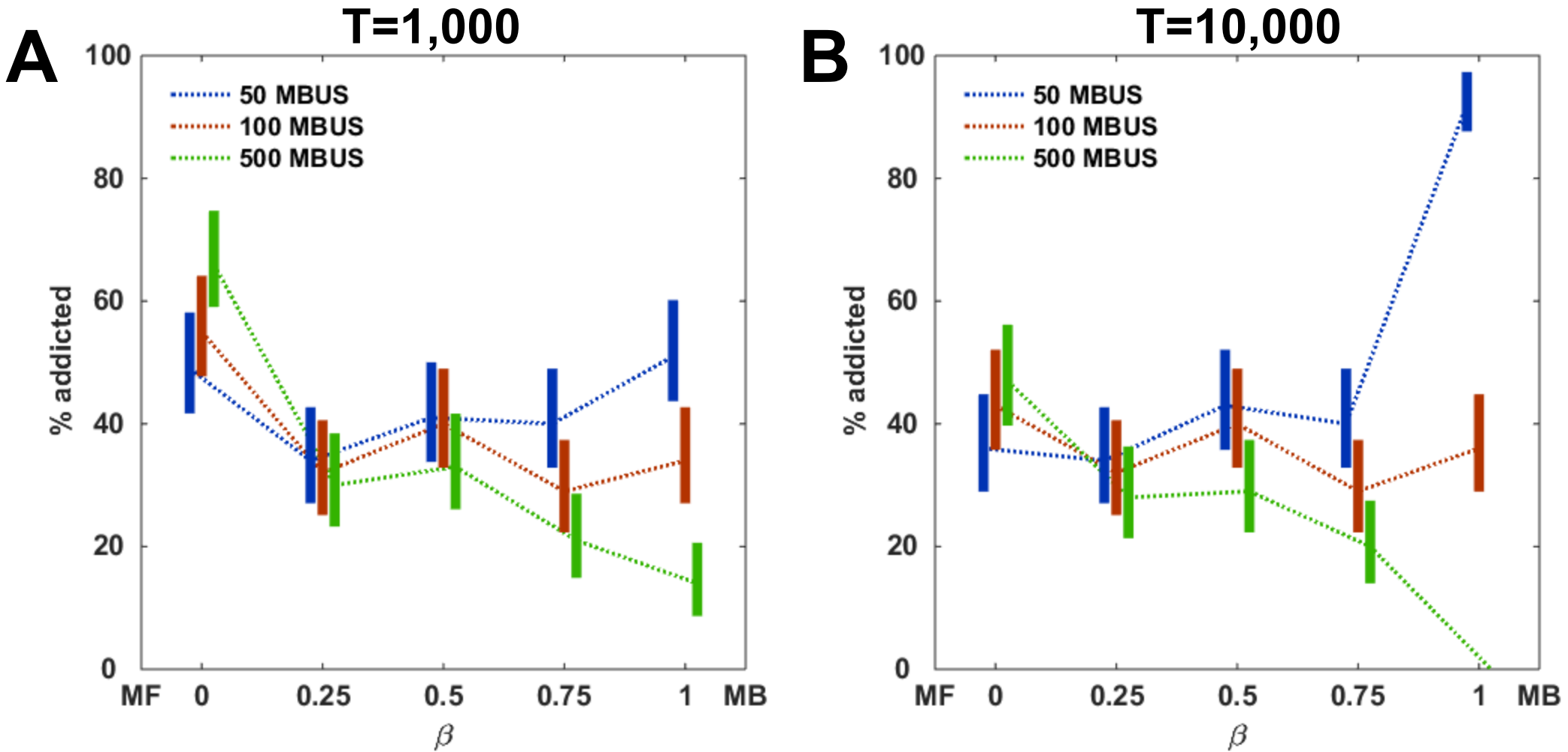

Dual Learning Agent
Q^{MF} Action Values
Model Free System
Q^{MX} Action Selector (β phenotype)
Environment Model
Model Based System
Planner
Q^{MB} Action Values
Environment
State & Reward
Action

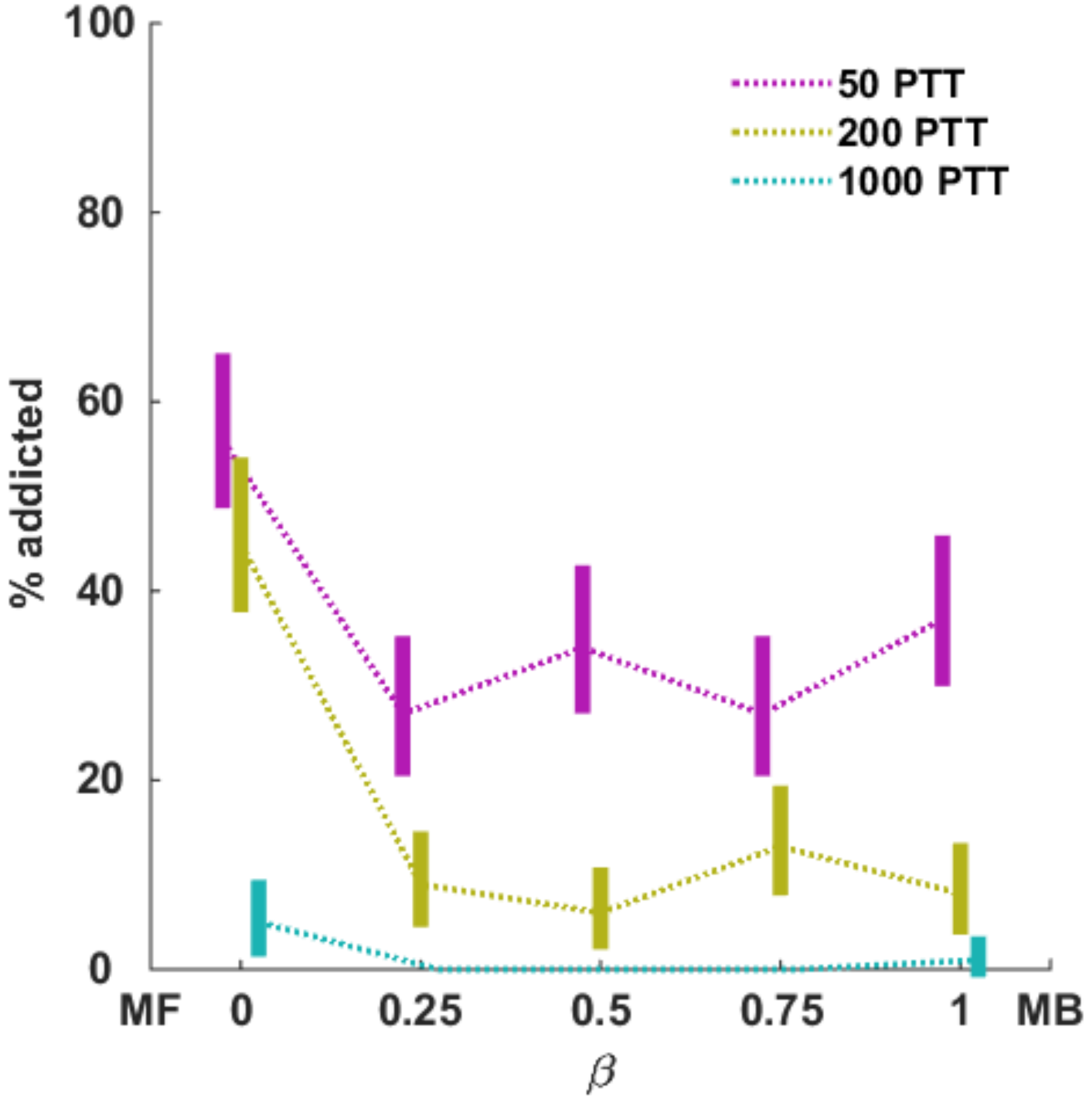

r=-4
r=-1.2
r=10
1
2
3
4
5
6
7
8
12
15
20
r=-1.2
r=1
r=-4
r=-4
Healthy Target
Neutral
Addictive Reward
Aftereffects
a^{s2-7}
a^{g}
a^{w}
a^{d}
P=1
P<1
P<<1

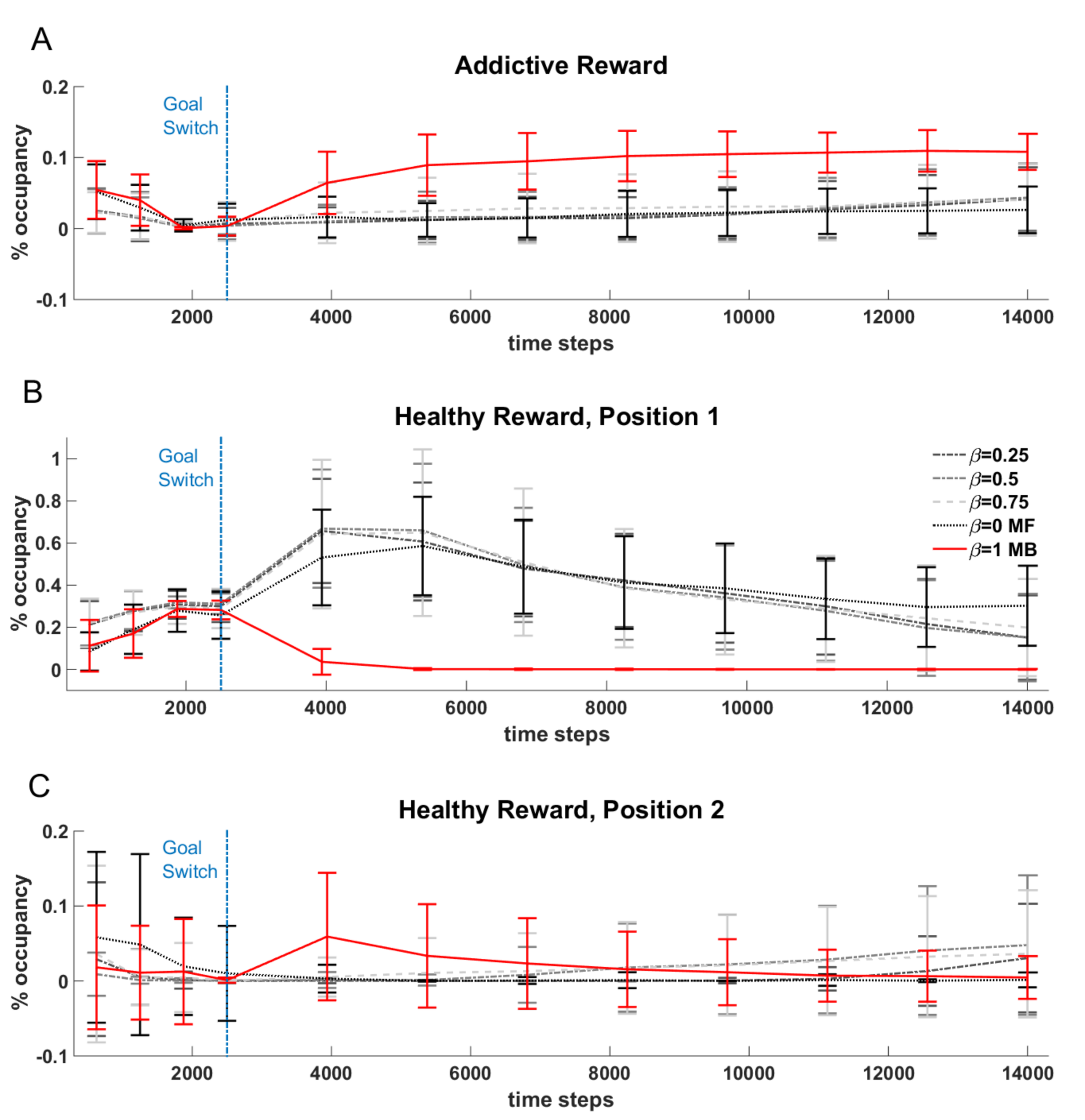

Purely Model Free β = 0
A
iInit
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
β = 0.2
B
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
β = 0.4
C
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
β = 0.6
D
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
β = 0.8
E
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
Purely Model Based β = 1
F
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps


% addicted
60
50
40
30
MF
0
0.2
0.4
0.6
0.8
1
MB
β


A
β = 0
init
addictive
condition
Addicted
Healthy
%addictive reward
1.5
1
0.5
0
-0.5
10^1
10^2
10^3
10^4
10^5
10^6
time steps
B
C
probability density
8
6
4
2
0
0
0.5
1
% addictive reward


A
T=1,000
B
T=10,000
50 MBUS
100 MBUS
500 MBUS
% addicted
100
80
60
40
20
0
MF
0
0.25
0.5
0.75
1
MB
β


100
80
60
40
20
0
% addicted
50 PTT
200 PTT
1000 PTT
MF
0
0.25
0.5
0.75
1
MB
β


A
Addictive Reward
Goal
Switch
% occupancy
0.2
0.1
0
-0.1
2000
4000
6000
8000
10000
12000
14000
time steps
B
Healthy Reward, Position 1
Goal
Switch
% occupancy
1
0.8
0.6
0.4
0.2
0
2000
4000
6000
8000
10000
12000
14000
time steps
β=0.25
β=0.5
β=0.75
β=0 MF
β=1 MB
C
Healthy Reward, Position 2
Goal
Switch
% occupancy
0.2
0.1
0
-0.1
2000
4000
6000
8000
10000
12000
14000
time steps