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

*dimitri.ognibene@essex.ac.uk

## Abstract

Several vulnerabilities have been identified as responsible for triggering addiction, or the repeated execution of actions despite their adverse consequences. In the present study we highlight a further vulnerability, proposing that the interaction between bounded rationality and environment complexity can increase the chances to develop addiction-like behaviours. Intuitively, resources and time required for deliberation, situation recognition, exploration or habit formation increase as a function of the depth and richness of details that must be considered for choice selection. Thus, multiple components and processes of Reinforcement Learning (RL) decision systems are either directly or indirectly affected by environment complexity. In a new RL model characterised by a dual-learning system, we show that functioning model-based and model-free selection systems are vulnerable to highly rewarding, but suboptimal, actions followed by several undesired (punishing) effects with sparse peaks of intensity, where the environment overloads both exploratory and computational capabilities of the agent. Our simulations show that the suboptimal decisions to repeatedly pursue the high reward, despite its negative effects (i.e. addiction-like behaviour), may stochastically emerge and then be 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 a vulnerability to addiction that has been so far neglected.

## 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]. 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 [3, 6, 7, 9]. 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, 11–15].

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 bio-chemical hijacking of the dopaminergic prediction error signal [3, 7, 16–20]. This substance induced habitual dominance nicely accounts for aspects of addiction such as inelastic behavior in the face of changes in the environment or even in presence of punishing outcomes following drug consumption [16, 21]. However, substance induced habitual dominance alone falls short in explaining several other findings associated with addictive behaviours [3, 18]. 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 [3, 18, 23, 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 [1, 2, 8] and internet gaming [10] are also considered part of the addictive behaviours, despite the lack of any pharmacological interference that would drive the aberrant habit formation [9].

These issues have been partially addressed by hypothesising the presence of vulnerabilities affecting the deliberative system [3]: (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 deliberative/planning system are not well understood [11, 27, 28].

Similar to Redish and colleagues' work postulating that addiction can emerge in environments pervaded by incomplete or inaccessible information [9], our study highlights that all-accessible, but non trivial, environments can demand excessive resources to be explored, represented or computed. In turn, failure to meet these demands can lead to addiction-like behaviors, independent of a pharmacological interaction. Most investigations in computational models of addiction have so far focused on environments characterised by the presence of easy to compute outcomes. This simplification has distanced the computational models from the clinical evidence, 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 [29–31]. 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 [34–38]. For example, Kakade [39, Chapter 8] presents several theorems proving that the amount of experience, e.g. number of interaction samples, necessary to achieve a behaviour with a certain minimum degree of performances grows faster than the product of the number of states and actions. Indeed, complex environments are heavily demanding in terms of exploration and training [37, 40, 41] so that action-outcome ramifications might not be well represented in an internal model, driving suboptimal choice selections and policies. Even when these

representations are sufficiently correct, complex representations in the internal model result in high demand of cognitive resources to assess the stored action-outcome ramifications and plan an optimal policy. Thus, suboptimal planning can result from a simple mismatch between environment representations and cognitive resources, factoring in temporary forms of cognitive impairments, as for instance due to anxiety or stress, known triggers in addiction disorders and relapse after treatment [42–45].

Our simulations show that the development of addictive behaviours may be supported by specific features of the environment structure, due to their effects on both habitual and deliberative processes [37, 40, 46, 47]. Specifically, under conditions in which difficult to evaluate long term negative outcomes follow a high reward, addictive behaviours emerge and persist for extended periods of time. To illustrate this process, we tested our agents under condition of undamaged decision system and in presence of a model-based component aiding the learning process in the model free (Q-Learning) component. 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.** Our simulated agents (Fig 1) are controlled by a hybrid (or dual) model system [18, 48–51] that 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 $\pi$.

MF algorithms such as Q-Learning [52, 53] and actor-critic architectures are usually employed to model habitual behaviours [3, 48, 54–57]. 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 component is employed to implement planning processes [11, 26, 27, 41] on the basis of an explicit representation, in an *internal model*, of action-state relationships and associated rewards, as experienced in the environment. 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.

A key issue of most dual models is that they assume an ideal MB process [50, 58], characterised by a complete knowledge of the environment and unlimited computational resources, which therefore always leads towards optimal choices. However, these assumptions conflict with several constraints, or bounds, that apply to any biological MB system [59–66], possibly hindering our ability to model biologically plausible healthy and dysfunctional behaviours (as e.g. in addiction [3, 18]). To avoid these limitations, in our simulations we have employed, jointly with a standard MF component based on the Q-Learning algorithm in tabular form [52], a MB component that represents only direct experience, and that relies on *bounded* [59, 60] computational resources 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 [67], 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 [68], which is commonly adopted in Artificial Intelligence for complex environments models where simulations are easier than computations. However, the Early Interrupted Stochastic Prioritized Sweeping algorithm employed here is more

computationally efficient for small environments [69]. With a limited number of updates it can provide results of more stable quality than few Monte-Carlo Tree Search samples.

**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

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, termed $\beta$, as follows:

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

As in previous constructs [70], six values (1, 0.8, 0.6, 0.4, 0.2, 0) have been used for this parameters to simulate different behavioural phenotypes, in a spectrum between purely model-based ($\beta$=1) and purely model-free ($\beta$=0) reinforcement learning. Finally, agents select the actions that maximize $Q^{MX}$ in 90% of their selections, in a standard stochastic $\epsilon$-greedy *persistent exploration* strategy [71], so preserving random exploration for the remaining 10% of selections.

**Environment.** We tested our hypothesis that suboptimal, addiction-like, behaviours can emerge without a pharmacological interference or a MB-MF malfunction, in an environment structure (Fig 2) that allows for both optimal and suboptimal non-trivial behaviours. For the case of suboptimal behaviour, a high immediate reward (i.e. the addictive reward as for instance in gambling) is followed by a series of inescapable and mixed-in-value negative outcomes or punishments (i.e. the multifaceted effects addictive behaviour has on the social life and health of the addicted).

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 healthy reward (moderate reward, $R_g = 1$) if they accessed the relative state and from there be brought back to the initial state. 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, $R_d = 10$), which was inescapably followed by the *after-effect segment*. 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 efter-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.

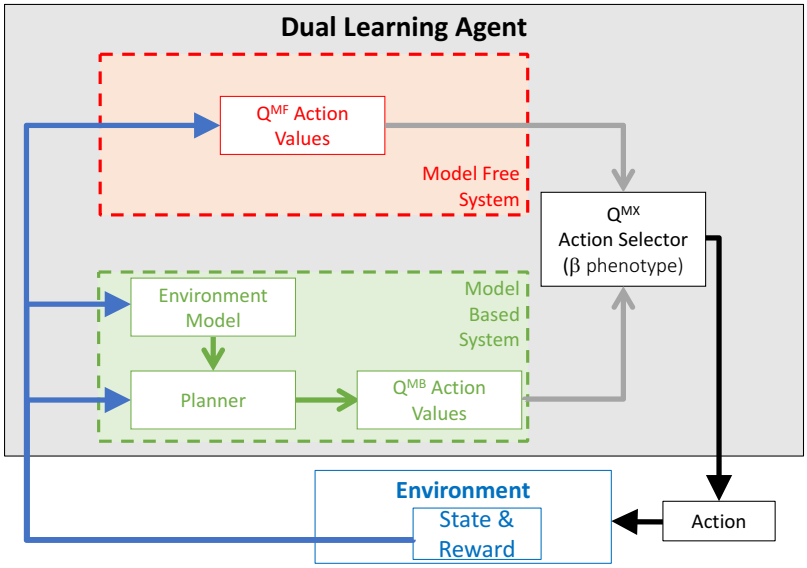



**Fig 1. Dual Learning Agent.** The decision making architecture includes: (i) a model free component (MF), which updates action values using value prediction error computations; and (ii) a model based component (MB), which generates an internal model of the environment, based on experienced action-outcomes and bounded computations. Action-outcome estimations derived from the two components are combined linearly according to a balance parameter, $\beta$, 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 behaviour trajectories for all populations show a significant bifurcation, where few initial choices strongly determine future choice preferences. Independent of differences in the parametrisations regulating MB/MF balance, agents seem to rapidly acquire a stable behaviour, marked by the near-exclusive preference towards either the healthy or the addictive state (Fig 3). The *healthy behaviour* is reached after less than 300 steps, across populations, and it is maintained for the entire time-length of the experiment.

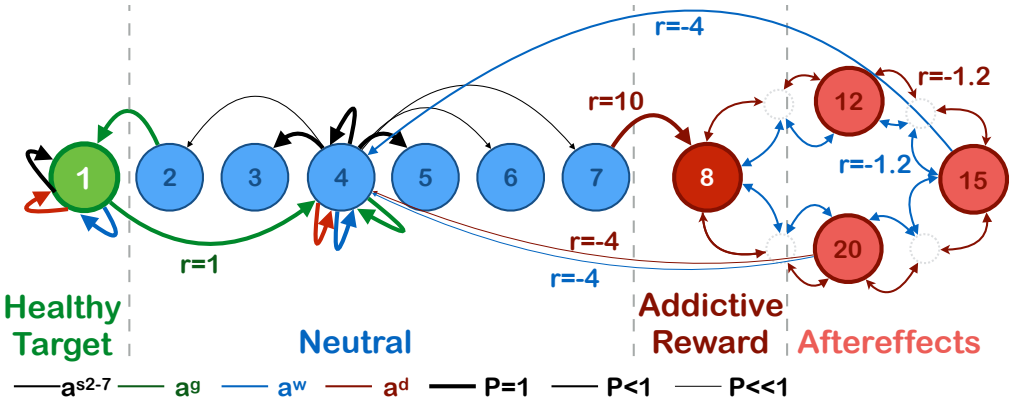



**Fig 2. Illustrative representation of the environment.** The states are disposed in a linear arrangement: on the left (number 1) a state associated with a healthy reward, on the right (number 8) a state associated with an addictive non-drug reward (e.g. gambling), separated by 6 neutral states that 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.

Conversely, *addiction trajectories* are 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 5), or more than 300 times the number of steps required by the healthy agents, with identical parametrization, to engage in the optimal behaviour (cf. conditions leading to extended suboptimal behaviour of the MF component are also discussed by [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 research on drug based addiction by adding a positive drug dependent term to the prediction error [16].

Confirming previous findings (cf. [70]), the balance of MB and MF components in the hybrid model significantly affected the percentage of agents in a population that developed a preference for the addictive reward. Higher resistance to addiction was demonstrated by populations characterised by *intermediate* values of $\beta = 0.4$ and $\beta = 0.6$, in comparison with those regulated by *extreme* values of $\beta = 0$, $\beta = 0.2$, $\beta = 0.8$ and $\beta = 1$ (see Fig 4).

To converge to optimum, when the model of the environment is known, the

**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 |
|---|---|---|
| $\alpha$ | MF learning factor | 0.05 |
| $\gamma$ | 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 |
| $\epsilon$ | Exploration Factor | 0.1 |

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 (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 is at maximum a small fraction of the 4K updates necessary for optimal estimation). 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. 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, and a fall to 0% in the number of addicted subjects, after 10K time steps. These behavioural trajectories result from the interaction between environment and cognitive resources. Any increase in the degree of complexity for the former results in an increased demand for the latter, 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

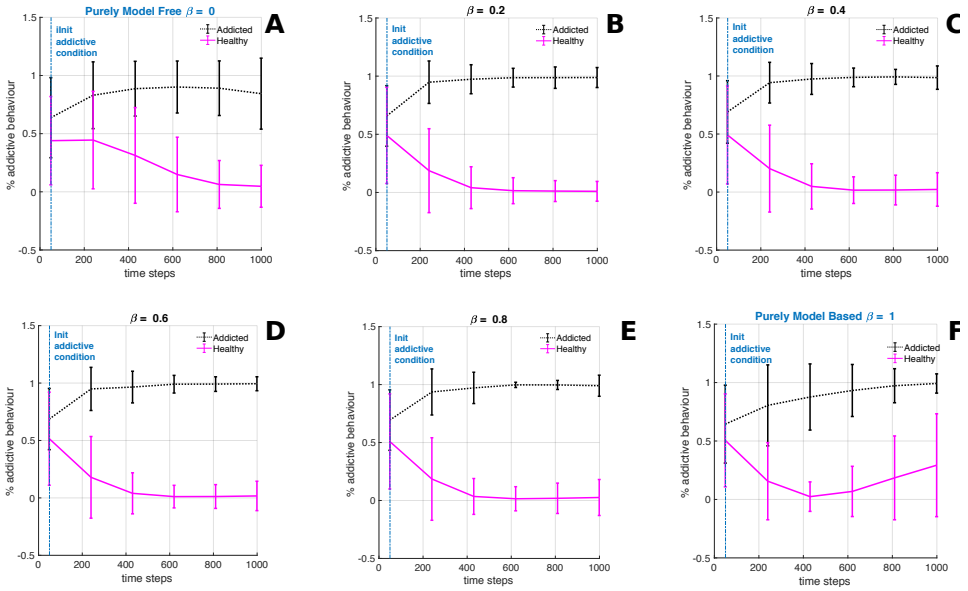



**Fig 3. Behavioural trajectories illustrating the ratio of healthy vs addictive reward state selections for addicted and healthy subjects.** The six panels highlight different behavioural trajectories, depending on $\beta$ values, which represent MB/MF balance per population. Addicted agents are defined 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 $\beta$ 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 preferences clearly emerges between addicted and healthy agents, for all parametrizations.

may require prohibitive and biologically implausible amounts of resources and exploration to replicate a result close to the described 500 MBUS population trajectory.

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 and therefore the execution of stochastic action selections. In turn, these initial choices shaped the initial value estimations, biasing future exploration depending on which rewarding state was found first by chance (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 acquisition 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 an $\epsilon$-greedy algorithm 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 $\beta$, and confirming the

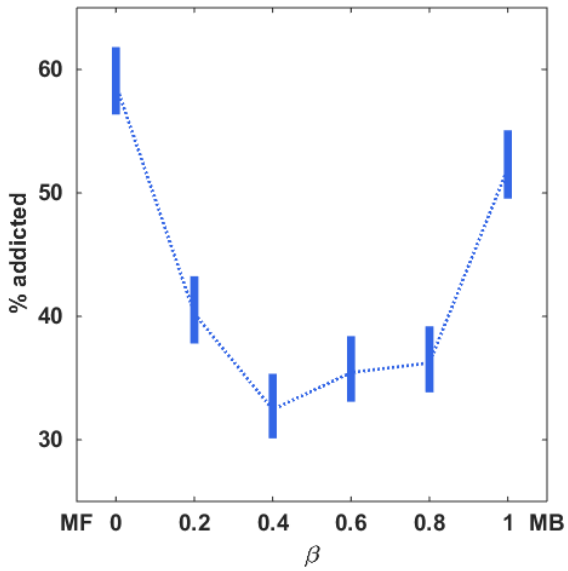



**Fig 4. Percentage and confidence intervals of addicted subjects per population, varying $\beta$.** Different $\beta$ values controlling the balance between MF and MB components were used for distinct populations of 900 simulated subjects. Addicted agents are those that during the observation period, 1000 time steps, acquire the addictive reward more often than healthy reward. The percentage of addicted agents per population varied as a function of $\beta$ values, where intermediate values showed a lower percentage of addicted agents (cf. [70]).Confidence interval were estimated assuming two-tail distribution and 95% confidence.

general higher resistance to addiction of the balanced MB-MF systems (intermediate values of $\beta$), in comparison with the unbalanced ones (extreme values of $\beta$).

Finally, we tested whether sudden environment changes could ignite addiction in agents that had developed the optimal healthy strategy [42, 45]. 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 significant 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 learning system of a biological agent. This phenomenon is mediated by hyper-physiologic phasic responses of dopamine to drug consumption, which in turn lead to non-compensable positive prediction error and unbounded growth of the perceived value of drug related actions. Thus, the consumption of sumbstances of abuse is assumed to trigger a

dopamine-mediated process of aberrant reinforcement, leading to habitual decision making, compulsive responses to drug-related stimuli and inelastic behaviour in the face adverse consequences [12, 72–74].

Despite significant advances in capturing important and complex features of addiction behaviour [11, 18, 19], this model remains primarily an expression of a malfunction of the MF component and therefore it leaves important questions unanswered [20, 75]. In particular, the role of the MB component in addiction is controversial. 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, non-habitual and non-pharmacological forms of addiction, which cannot be explained in terms of drug-induced hijacking of the prediction-error mechanism, seem to suggest that the MB component [11] might drive addiction-like suboptimal behaviours, independent of MF malfunctions.

In this study, we have focused on the second of these two key issues, investigating the role of the MB component in driving addiction, independent of the alterations that drug consumption or addictive behaviours may have on the decision components themselves. To this end, we have simulated the emergence of MB-driven addiction-like suboptimal behaviours, under conditions deprived of the simulation of drug consumption. We have found that a relatively complex environment, explored by an agent characterised by a bounded MB component, resulted in a failure in correctly estimating positive and negative outcomes associated with the available courses of action. Our simulations show that, when a segment of the environment is characterised by the presence of an immediate reward and more difficult to learn negative outcomes, an agent bounded by limited exploration or cognitive resources can mistakenly chase obvious rewards and ignore the associated, non-obvious, ramifications of punishments. Under these conditions, the simulated agents mimic non-habitual, addiction-like, inelastic behaviours in the face of adverse consequences, despite the absence of a classic form of drug-induced aberrant prediction error signal or the associated over-learning (as e.g. in gambling or videogaming).

Contrarily to the role ascribed to the MF component by the classic RL approach, our simulations point towards a more complex role played by the MB in addiction. We described how high computational resources and deliberative capacities were able to accurately represent and assess the environment chosen for our tests, thus greatly reducing the chances to develop addiction. This finding is consistent with previous studies indicating reduced contribution of the MB component may be a risk factor for addiction [76]. However, the exponential growth of complexity that would be associated by the use of more ecologically plausible set-ups 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), stronger influence of MB component (small values of $\beta$) or in the availability of computational resources (e.g. due to stress, a known trigger for addiction [43, 60, 64, 65, 77–79]). This apparently contradictory role ascribed to the MB component in addiction is consistent with multiple studies that have highlighted how brain areas commonly associated with MB decision making can either increase or decrease their activity in addict individuals, under different tasks and conditions [80–84].

The problem of environment complexity, with the associated functional shortfalls caused by exploration costs, affects multiple components of the decision system, including MF, attention control, Pavlovian, and state identification [3, 9]. Nonetheless, the MB system engages in complex deliberations, policy assessments and goal selections

that can lead to ignore or miscalculate the extent of the adverse consequences associated with pursuing addictive behaviours. This MB vulnerability can complement and interact with previously described ones [3]. In pharmacological forms of addiction, erroneous representations and assessments of the environment can aggravate the behavioural symptoms associated with the classic drug-induced MF malfunction and caused by non-compensable prediction errors. This interaction can account for those complex non-habitual drug-seeking behaviours that are not triggered by the presence of drug-related stimuli [3, 18, 23, 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 (e.g. after a successful treatment has resulted in extinction), the MB component may keep pursuing sub-optimal policies, contributing to both craving and relapse [85]. Finally, the vulnerability we have described can be seen as ideally contiguous with those associated with state identification errors [9, 86–88]. 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 states where addictive behaviours would always have high values [9, 89]. 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 resources availability allowed us to account for the presence of different behavioural trajectories or phenotypes. We have observed behavioural differences arising from changes in key parameters (e.g. depending on the amount of cognitive resources available to the agent), as well as unpredictable behavioural bifurcations which depended on less than *few hundreds* initial stochastic exploration steps and lasted up to *hundreds of thousands* of 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 model and results have interesting implications for treatment development. 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, 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 association between normally occurring negative outcomes and the addictive behaviour. 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 [90, 91] and stress [45, 92]) 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 simulation of complex environments with features that are designed to resemble a more ecologically plausible form of addiction, with a prominent role for the MB component. 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 in the future.

## 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. Daw ND, Dayan P. The algorithmic anatomy of model-based evaluation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1655):20130478.
12. 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.
13. Norman DA, Shallice T. Attention to action: Willed and automatic control of behavior. Cognitive neuroscience: A reader. 2000; p. 376–390.
14. Nadel L. Multiple memory systems: What and why, an update. Memory systems. 1994;1994:39–63.

15. 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.

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

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

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

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

20. 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.

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

22. 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`.

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

24. 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.

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

26. 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.

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

28. 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.

29. 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.

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

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

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

33. 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.

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

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

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

37. 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.

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

39. Kakade SM. On the sample complexity of reinforcement learning. University of London; 2003.

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

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

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

43. 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.

44. 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.

45. 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.

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

47. 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.

48. 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.

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

50. 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.

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

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

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

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

55. 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.

56. 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.

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

58. 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;.

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

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

61. 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.

62. 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.

63. 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.

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

65. 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.

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

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

68. Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling P, Rohlfshagen P, et al. A survey of monte carlo tree search methods. Computational Intelligence and AI in Games, IEEE Transactions on. 2012;4(1):1–43.

69. Domshlak C, Feldman Z. To UCT, or not to UCT?(Position Paper). In: Sixth Annual Symposium on Combinatorial Search; 2013.

70. Fiore VG, Ognibene D, Adinoff B, Gu X. A Multilevel Computational Characterization of Endophenotypes in Addiction. eNeuro. 2018;.

71. 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.

72. 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.

73. 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.

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

75. 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;.

76. 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.

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

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

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

80. 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.

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

82. 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.

83. 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.

84. 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.

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

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

87. 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.

88. 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`.

89. 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`.

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

91. 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.

92. 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.

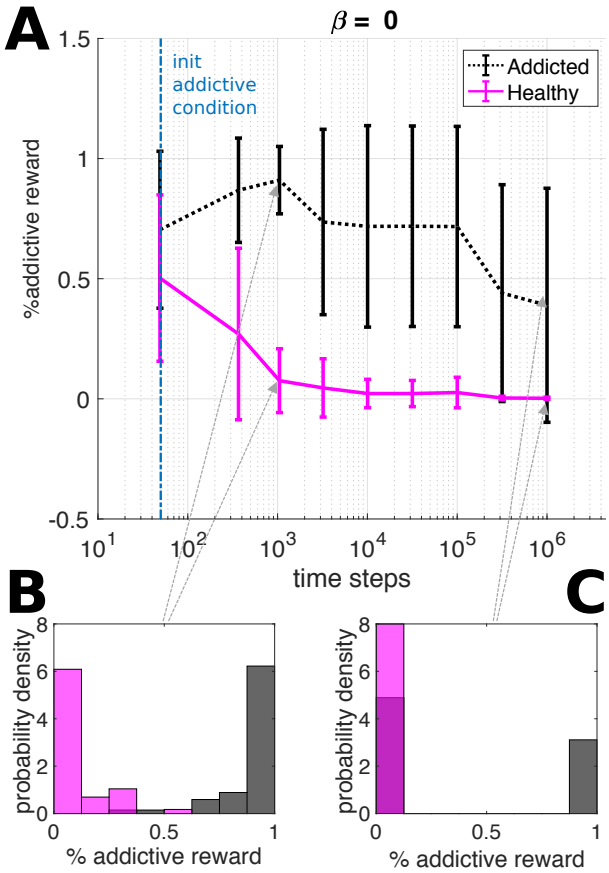



**Fig 5. Long runs with logarithmic time scale.** Behaviour expressed by purely MF agents ($\beta = 0$) was recorded and averaged over 100 runs, separating the addicted agents (high preference for the addictive reward state in the first 10,000 steps) from the healthy ones (the remaining agents, which 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).

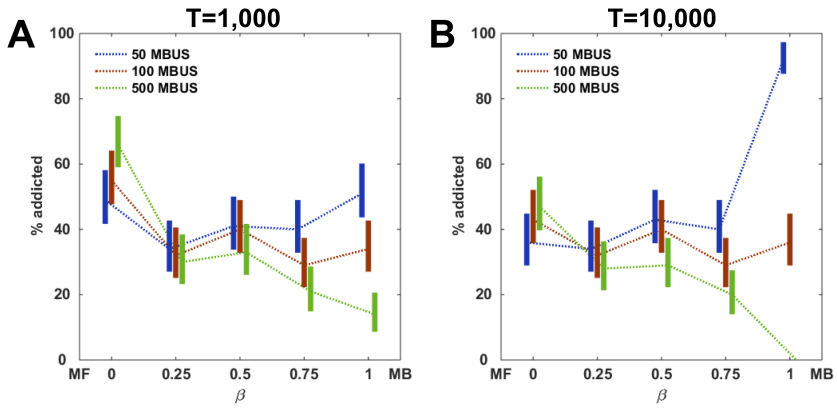



**Fig 6. Preference ratios and confidence intervals of agents expressing addiction-like behaviour within each parametrization of cognitive resource bounds (Model Based Updates per Step [MBUS]) and MB/MF balance factor $\beta$.** Initial performance (panel A, analysis on the behaviours in the interval 900 to 1000 timesteps) shows a significant preference for the selection of the addictive reward state, across all values of $\beta$ 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 $\beta$ 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 $\beta$, 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 $\beta$ 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 $\beta$), 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 $\beta$ value.

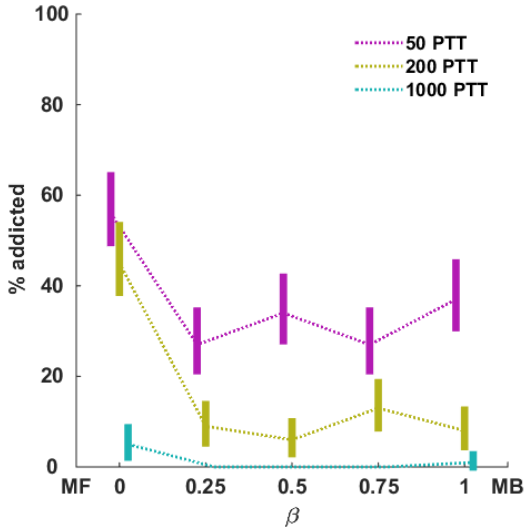



**Fig 7. Changes in behavioural trajectories as a function of pre-training time (PTT, timesteps in safe phase) and $\beta$ parameter (MB/MF balance).** Exposure to the environment before the introduction of the addictive reward decreased the probability of addiction across all sets of parameters or populations. 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.

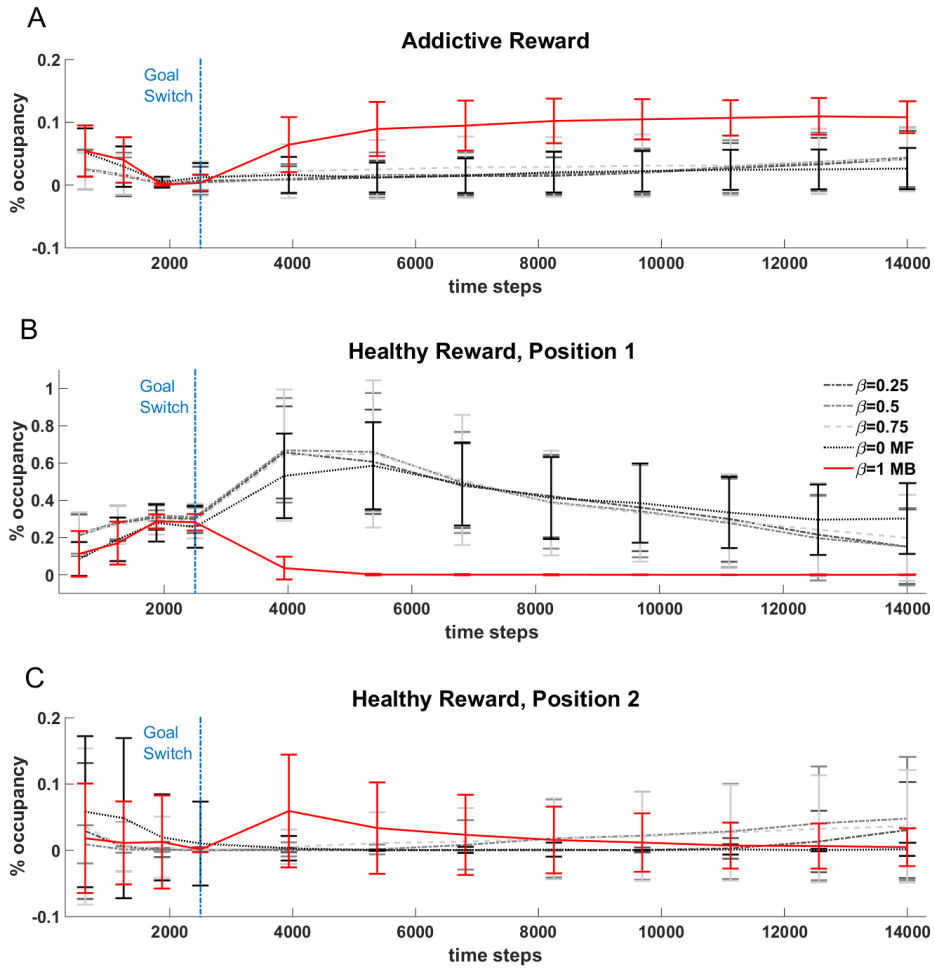
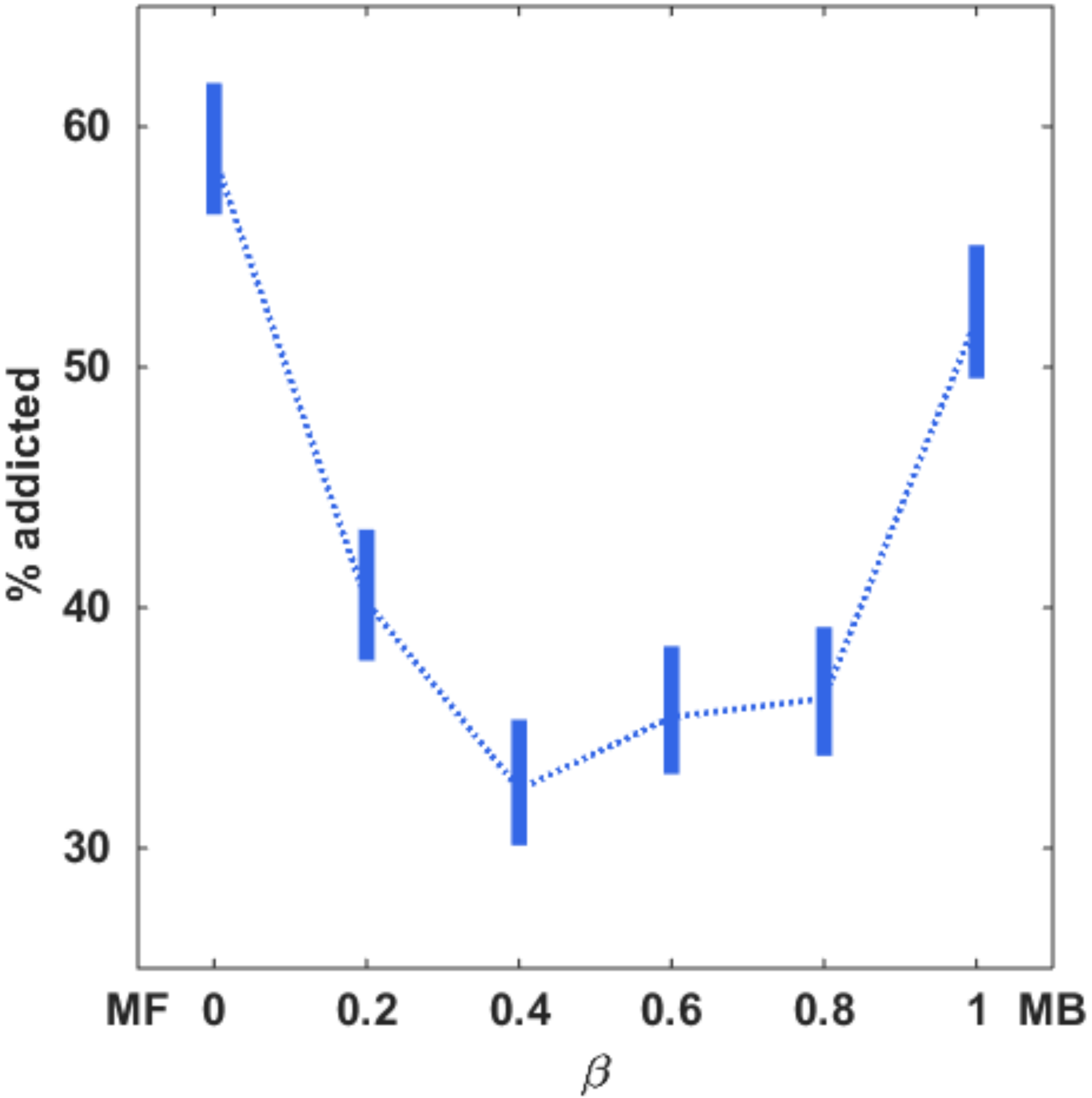
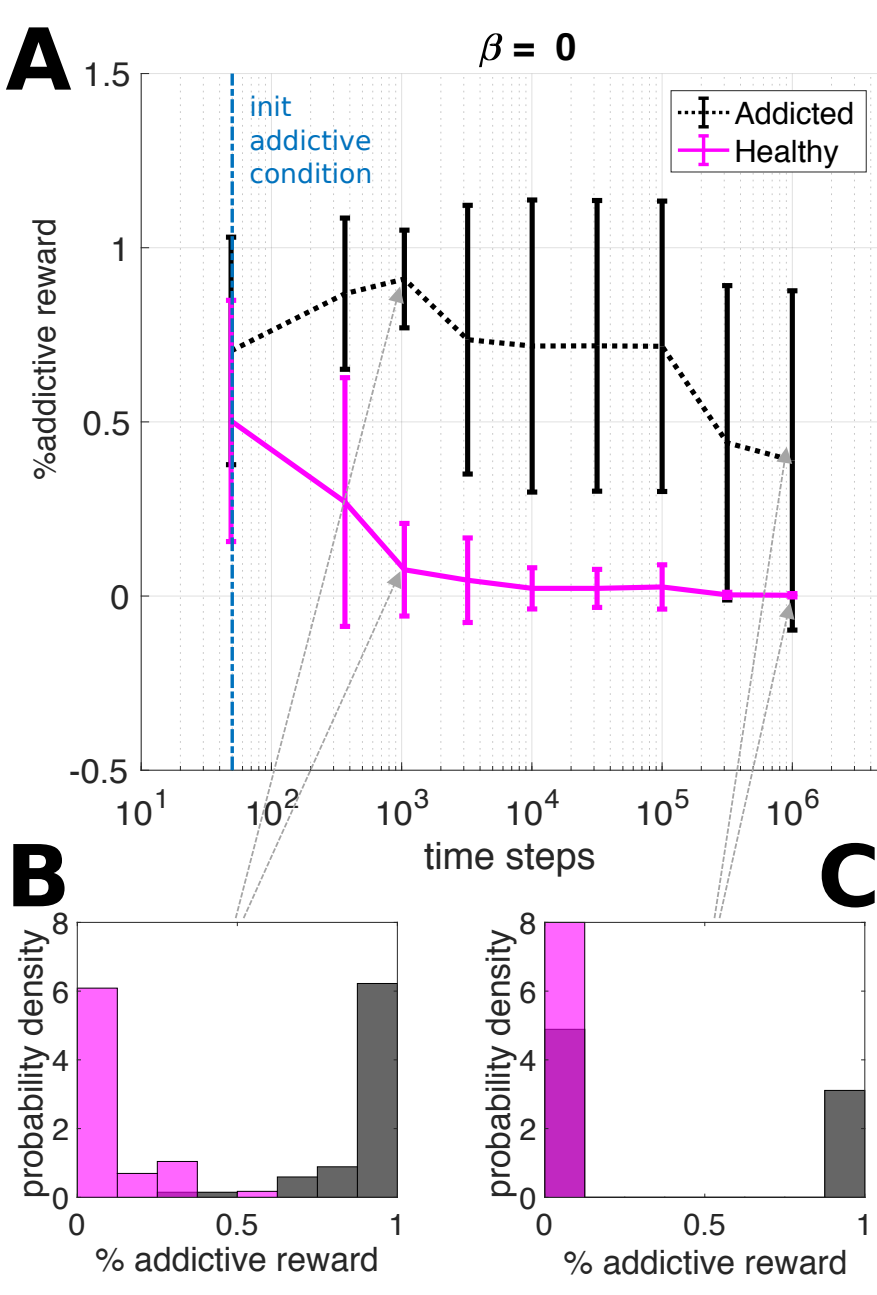
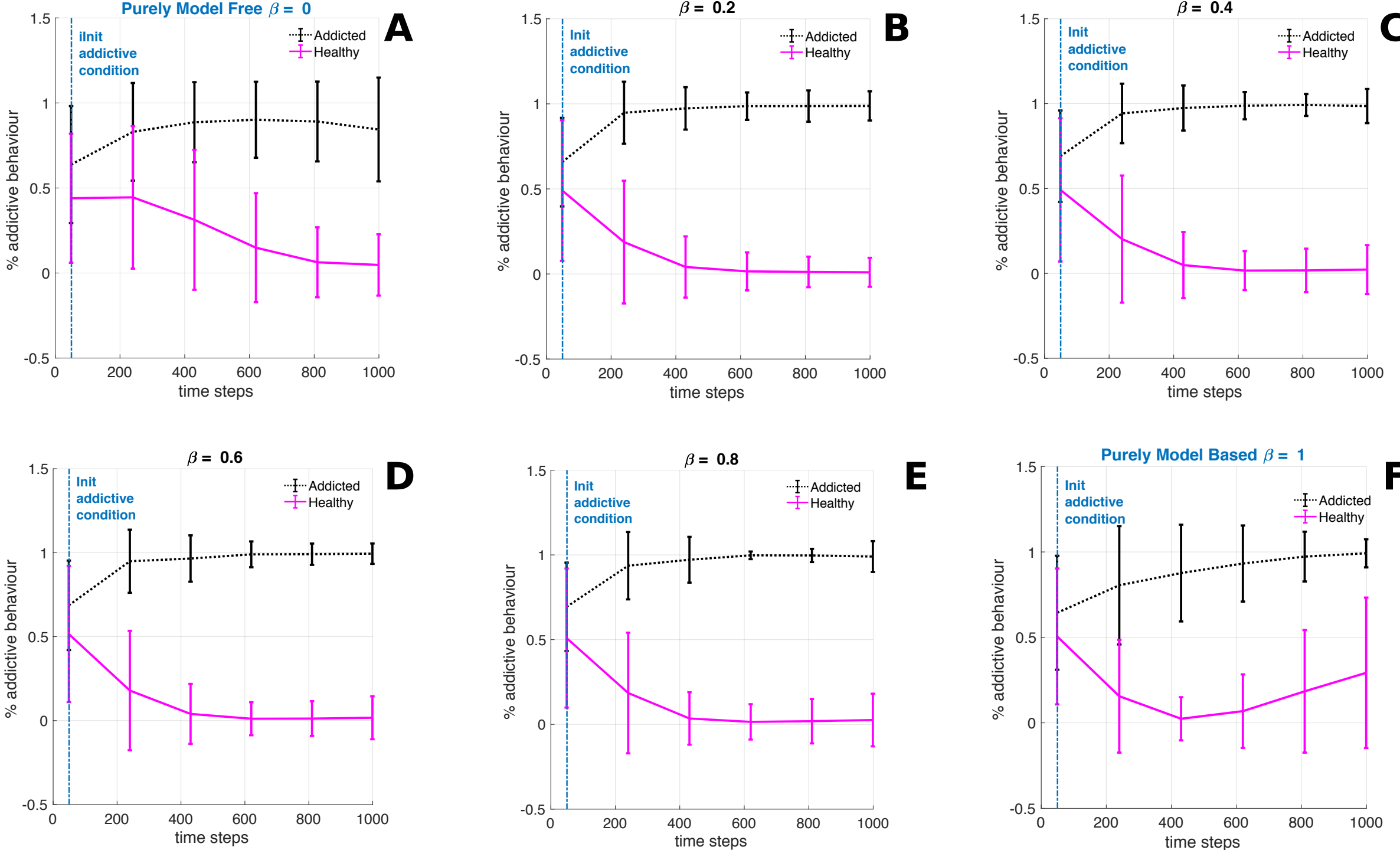
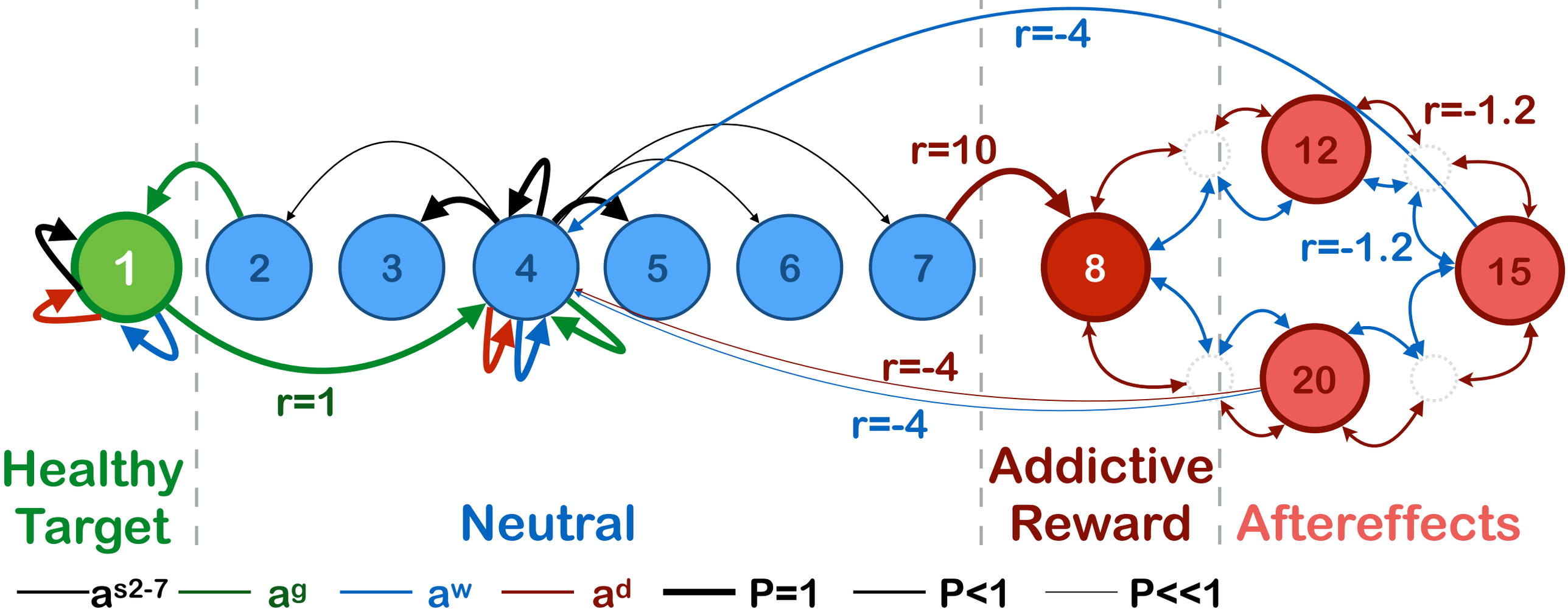


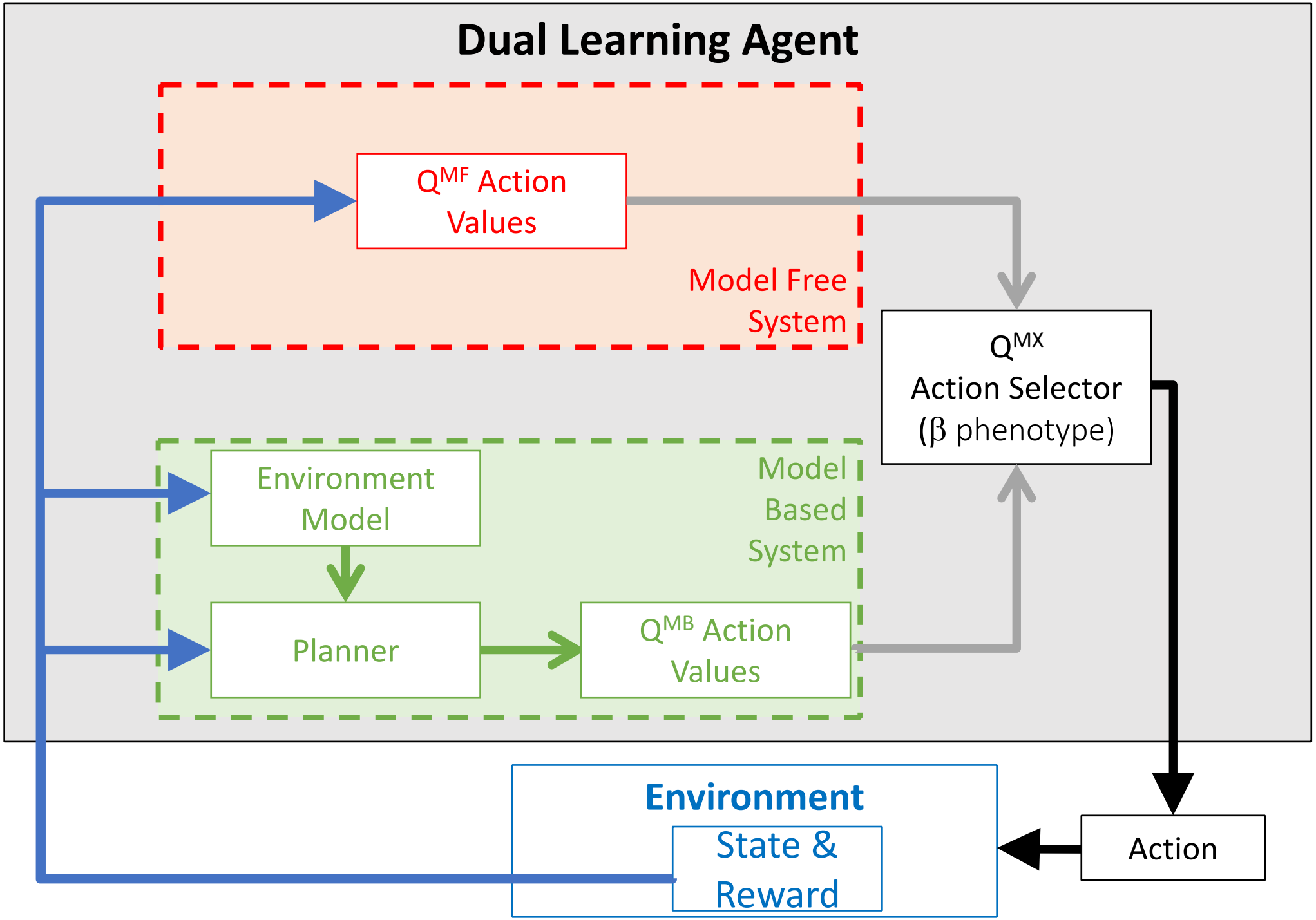

**Fig 8. Effects of environment change on healthy subjects** This figure illustrates the effects on behavioural trajectories caused by a change affecting the position of the healthy reward state, depending on the parameter $\beta$ regulating MB-MF balance. The change takes place at time step 2500, when 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.

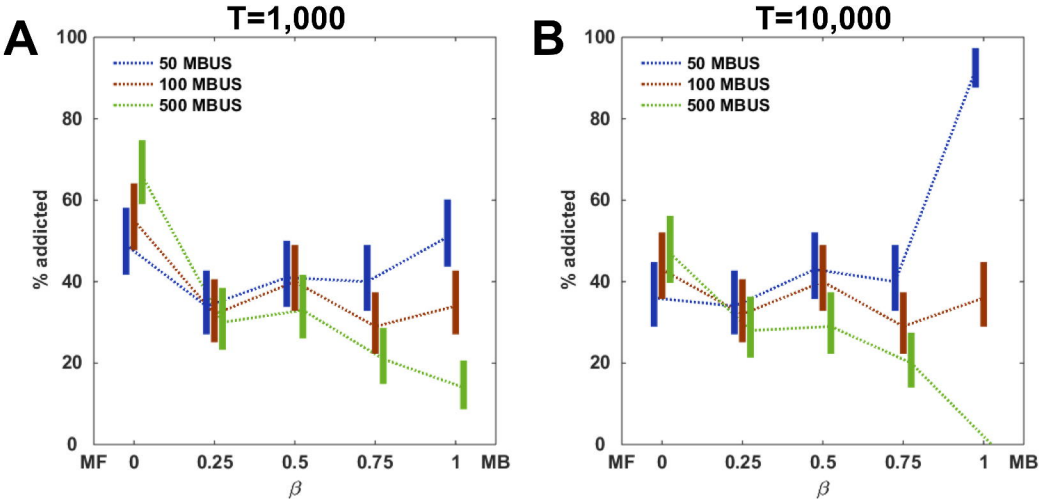

% addicted
60
50
40
30
MF
0
0.2
0.4
0.6
0.8
1
MB
β

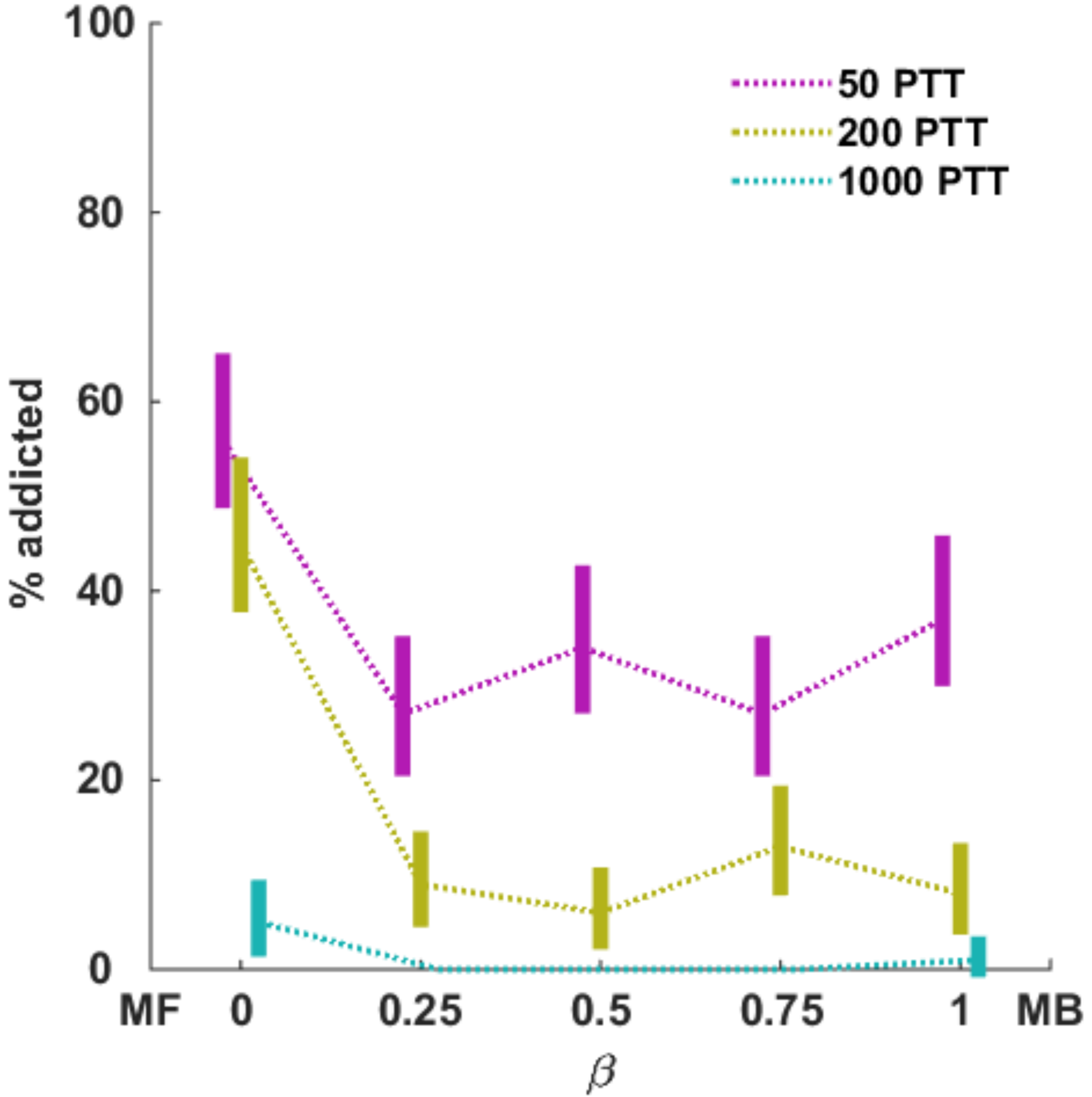

A
β = 0
init
addictive
condition
Addicted
Healthy
%addictive reward
1.5
1
0.5
0
-0.5
10^1
10^2
10^3
10^4
10^5
10^6
time steps
B
C
probability density
8
6
4
2
0
0
0.5
1
% addictive reward

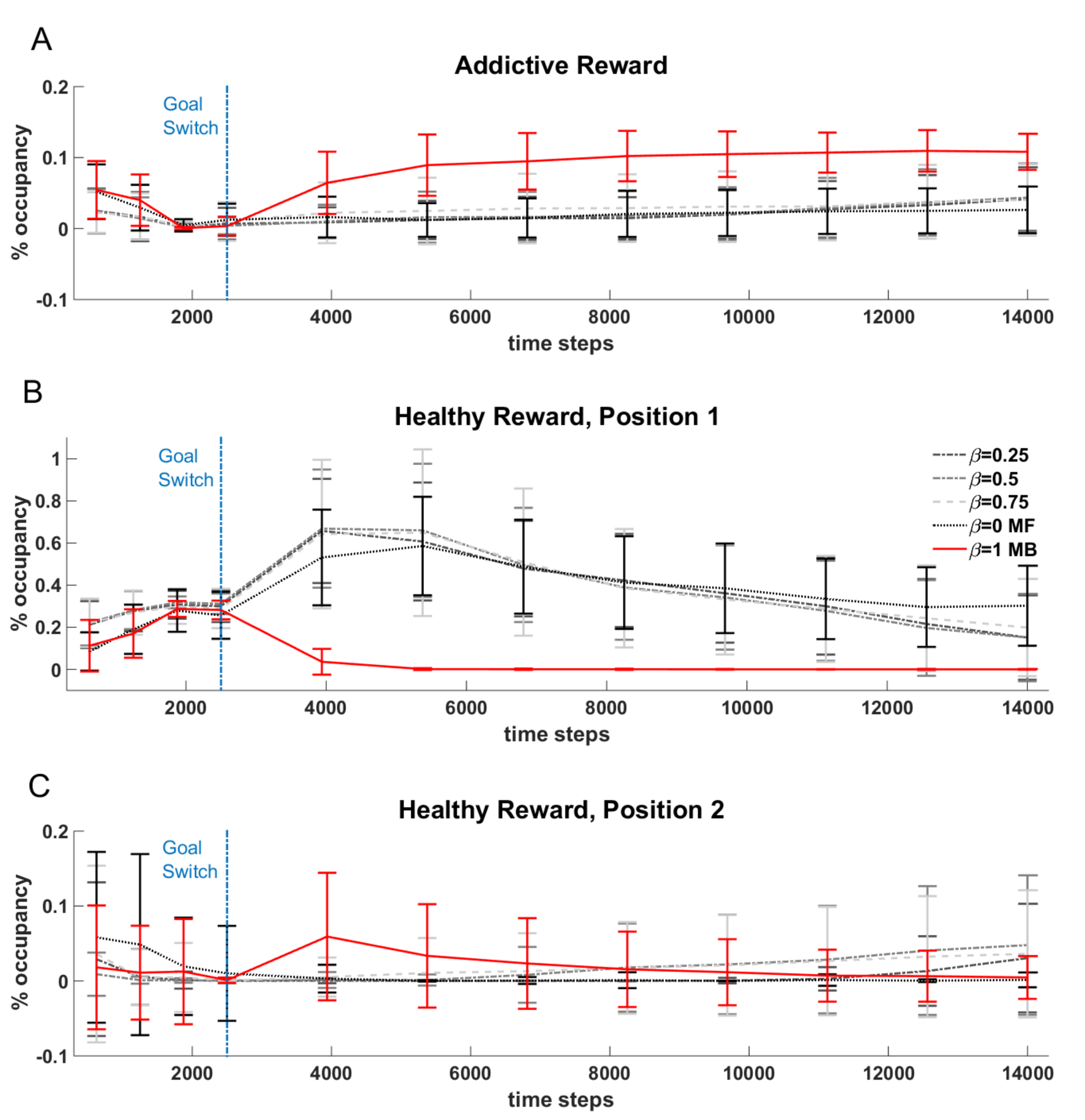

Purely Model Free $\beta$ = 0
A
iInit
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
$\beta$ = 0.2
B
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
$\beta$ = 0.4
C
Init
addictive
condition
Addicted
Healthy
% addictive behaviour
time steps
$\beta$ = 0.6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


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


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


A
T=1,000
B
T=10,000
50 MBUS
100 MBUS
500 MBUS
% addicted
0
20
40
60
80
100
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