Addiction in a Bounded Rational Model: Role of Environment and its Exploration

Dimitri Ognibene[1], Vincenzo Guido Fiore[2], Xiaosi Gu[2]

1 Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
2 Center for BrainHealth, University of Texas at Dallas, Dallas, TX, USA

* dimitri.ognibene@upf.edu

Abstract

Existing reinforcement learning (RL) theories of addiction often rely on the 'dual-system theory of decision making'. This theory describes behaviour as a product of the interaction between a reactive, slowly adapting, model-free (MF) component, and a planning-oriented, flexible, and computationally demanding, model-based (MB), one. In addiction, the MF component is assumed to overtake the MB component, causing a behavioural shift towards aberrant reactive responses to drugs. RL theories ascribe this behavioural shift to substance induced bio-chemical effects, which are thought to hijack the standard temporal difference learning mechanism regulating the MF component. However, MF processes seem to account mainly for simple, repetitive and compulsive behaviours so that existing RL models have limited power in explaining: 1) the heterogeneity of drug seeking behaviours; 2) the different mechamisms of actions of the different drugs; 3) addictive behaviours which are not associated with drug intake (e.g. gambling). In this paper, we propose that a non-simplistic experimental environment may account for such multifaceted phenomena in addiction, suggesting that heterogeneous addictive behaviours can arise also without modifications or malfunctions of the standard dual system. Environment complexity is often neglected in laboratory setups, despite its role in modulating the adaptation of both MF and MB components. In particular, environment complexity directly affects the MB decision process due to bounded computational resources, and indirectly affects both the MB and MF learning processes, as exploration demands increase as a function of environment complexity. Thus, we employed an environment where highly rewarding actions are followed by several sparse undesired (punishing) effects, in a pattern that exceeds 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 parameters regulating the interaction between MF and MB control modalities. Our simulations suggest that addiction may also arise independently of a pharmacological component (e.g. gambling and videogaming) under specific conditions of environment complexity.

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 factor leading to addictive behaviours. Addiction may emerge from the dynamic process of adaptation in complex environments. These computational findings highlight the importance of the environment, learning dynamics, and the computational capabilities of an agent in the development of disorders, suggesting this 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], and other putative forms of behavioural addictions, such as internet gaming [9].

Several theories and computational models have been proposed to explain the repetition of suboptimal decisions typical of addiction [3, 6, 7]. In most of the these theories, decision making is assumed to result from the interaction of multiple systems relying on different learning and computing principles. This composed structure leads to complex interactions and interface problems that have been seen as candidate causes of several possible vulnerabilities, which are affected by the interference of drug-induced reinforcements [3, 10–14].

A classical distinction defines decision systems as controlled by either habitual or deliberative/planning systems [12, 15–18]. Habitual behaviour consists of stimulus-response associations which are consolidated or reinforced by unexpected outcomes or rewards (Thorndike's law of effect [15]). This behavioural modality is characterised by slow flexibility, but high cognitive efficiency in terms of time and resources it requires. In contrast, planning allows flexible decision making by assessing the outcomes allowed by the environment, among the accessible courses of action. This behavioural modality requires an 'internal model' of the agent-environment interactions, to link candidate actions to their expected outcomes, on the grounds of previous experience. A complex environment requires a substantial amount of exploration [19–21] to be correctly represented in a useful and precise model. In turn, the assessment of the ramification of action-outcome 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 [22–24].

Despite its importance for both planning and habitual responses, the relation between behaviour and environment complexity has received limited attention in computational studies of addiction. It is generally accepted that environment complexity hinders cognitive performance and may lead to maladaptive decisions beside addictive behaviours [25–30]. However most investigations, both computational and animal models, have focused on simplified environment to ease the analysis of the neural correlates of the expressed behaviour. 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 [31–33].

Current prominent theories of drug addiction build on the idea that abnormally strong drug-seeking habits come to dominate behaviour due to drug induced bio-chemical hijacking of the reward estimation process [3, 7, 34–37]. Disrupted estimations may lead to aberrant habits through several mechanisms, such as inhibition

of the planning system, over-evaluation and reinforcement of drug-related actions [38–40] or increased discount of longer term (adverse) outcomes [3, 7, 41]. However, by focusing on substance induced habitual dominance, current research falls short in explaining several findings associated with addictive behaviours [3, 35]. 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)" [42]. Such temporally extended activities, are often novel, complex and context-dependent [3, 35, 43, 44], and are unlikely to be justified by appealing solely on habitual processes and stimulus-response conditioning. In a similar manner, substance induced habitual dominance view also fails in explaining phenomena such as craving, which can occur even without exposure to conditioned stimuli (but see [45, 46]). Second, several experiments have shown addiction results in hypersensitivity in brain areas associated with planning processes, during both non-drug and drug related reward encoding [47–49]. Third, not only different drugs present very different modes of action but addictive behaviours are described also for gambling [1, 2, 8] and internet gaming [9], which cannot rely on the hypothesis of the pharmacological hijacking of habit generation.

Several drug related vulnerabilities have also been proposed for the deliberative process [3]: (i) misinterpreting a drug as beneficial or useful for the agent; (ii) over evaluating the *immediate* reward connected to a drug ; (iii) over evaluating *sequences* of actions comprising drug intake and its consequences. However, the neural and computational mechanisms by which drugs can induce these effects on the planning system are far from being understood (but see [10, 50, 51]). In particular, it is debated whether dopamine, an ubiquitous target of drugs of abuse, can play a role in altering the planning system by affecting reward learning, in a way that would resemble the role dopamine plays for habits formation [10, 35, 52–54]. We will elaborate more on this in the next section.

In this study, we propose that a solution for these shortcomings can be found by focusing on two closely related factors that have been mostly neglected in addiction modelling: environment complexity and exploration. Our simulations show that the development of addictive behaviours may in some cases be supported by specific features of the environment structure due to its effects on deliberative processes and exploration, as decision making relies on the knowledge acquired in the environment [19, 20, 55, 56]. Specifically, when 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.

Modelling Addiction in the Reinforcement Learning Framework

Introduction of Reinforcement Learning concepts

The model we propose is grounded in the Reinforcement Learning (RL) framework [57, 58]. For the readers who are not familiar with the RL framework this section provides a short introduction to its main concepts. Readers familiar with this literature may wish to directly proceed to the following sections.

While RL has been developed within the field of Artificial Intelligence (AI), it is widely adopted in cognitive research [10, 39, 50, 59–62] because it allows to quantitatively capture several important features of biological behaviour, such as learning processes and sequential decision making.

In RL computational theory, agents learn and adapt their behaviour to define a strategy of action, or policy π, that maximises *value*. Value is a quantity that combines short and long term average (or 'expected') *rewards* (or positive outcomes) and *punishments* (or negative outcomes) when a specific strategy of action is followed. The value associated with a specific environmental condition (state) s_t and agent response a_t is defined as:

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \mathrm{E}[\sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i} = \pi(s_{t+i}))|s_t, a_t] \quad (1)$$

In Eq 1, $r(s, a)$ denotes the instantaneous reward, if positive, or punishment, if negative[1], received when action a is performed in state s. γ is a discount factor, comprised between 0 and 1, which defines the trade-off between immediate and long term rewards. The value of a state given the policy is defined as $V^{\pi}(s) = \max_a[Q^{\pi}(s, a)]$. For each environment there is an optimal policy $\pi^*(s) = \arg\max_a Q^{\pi^*}(s, a)$, which maximizes the value for every state [57]

In the RL framework, the distinction between habitual and planning processes of decision making has been classically reframed in terms of the separation between model-free (MF) and model-based (MB) classes of RL algorithms [10, 17, 63].

Habitual processes and resulting habitual behaviours are usually modelled ([3, 17, 59, 64–66] but see [67, 68]) with MF algorithms, such as Q-Learning [57, 69] and actor critic architectures. These algorithms maintain an estimation of the value of candidate actions $\tilde{Q}^{MF}(s, a)$ updated after the execution of each action on the basis of the obtained rewards and observed states. The process of updating is quantified with a variable δ_t, termed *reward prediction error*. To compute the prediction error, the algorithm considers any reward resulting from a selected action, the value of the state accessed by performing the action and the value of the state at the time the action was executed.

In Q-Learning updates the prediction error is defined as:

$$\delta_t = R(s_t, a_t) + \gamma \max_{a'} \tilde{Q}^{MF}(s_{t+1}, a') - \tilde{Q}^{MF}(s_t, a_t) \quad (2)$$

Value estimations are update according to:

$$\tilde{Q}^{MF}(s_t, a_t)_{new} = \tilde{Q}^{MF}(s_t, a_t)_{old} + \alpha\delta_t \quad (3)$$

where α is a learning factor comprised between 0 and 1.

The higher the value associated with an action, the higher the probability to execute it, thus defining the formation of a stimulus-response habit (in RL terms also known as model-free, cached or precomputed, policy). Intuitively, when in a certain context the execution of an action is followed by a reward and/or by a high value context, the value associated with the executed action increases reinforcing its probability of execution in the original context.

In contrast, planning processes are implemented by MB (or indirect) RL [10, 21, 46, 50] algorithms. These maintain an explicit *internal model* of the causal relationships experienced in the environment, including information about how a state changes following the execution of an action (state transitions distributions $p(s_{t+1} = s|s_t, a_t)$), with the associated reward ($\mathrm{E}(r|s, a) = \mathrm{R}(s, a)$). This internal model is then used for the *online* evaluation of actions consequences, achieved by internally iterating variations of eq.1 such as the Bellman equation [57], over multiple states and actions :

$$Q^{\pi^*}(s, a) = R(s, a) + \gamma \sum_s [p(s_{t+1} = s'|s_t = s, a_t = a) \max_{a'} Q^{\pi^*}(s', a')] \quad (4)$$

[1]Following RL terminology, simply 'reward' will be used when the sign is not relevant.

Both MB and MF processes ideally produce an optimal 'policy', or action strategy, under the assumptions of unbounded cognitive resources and infinite exploration [69].

To describe the interactions between the model-based and model-free processes several models have been proposed that have been termed *dual* or hybrid [10, 17, 59, 65, 70–73]

Biological implementation of RL and a role of dopamine Considerable evidence supports the idea that MB and MF processes are implemented by mostly distinct neural circuits [10, 11, 17, 74] These normally concur in driving behaviour [65, 70, 71, 73, 75], but they can even drive it independently of each other in special conditions, e.g. lesions [13, 14].

Imaging and recording data support the idea that the value prediction-error signal δ_t (Eq 2) that is used to train the MF component is implemented in the midbrain dopamine system [11, 76–78], which is known to be a target of many drugs of abuse [3]. However, also other areas correlate with the value representation including prefrontal cortex, striatum, and parietal regions [79].

The details of the neural implementation of the model-based RL process are less known. Current evidence support the hypothesis that the model-based neural implementation involves diverse cortical regions, the hippocampus and other subcortical structures [10, 16, 50, 51, 80].

It is quite unclear if value prediction error signals are involved in reward representation in the MB process [10, 35] learning. In fact, while MB influences have been found on the prediction error δ_t (Eq 2) used to train the MF component [66], the prediction error is a quantity connected to the value (Q^π in Eq 1) and as such it is implicitly dependent on the discount, the future sequence of states, and the policy adopted. Thus the prediction error does not suit the estimation of the MB reward associated with the current state which is instead independent from the policy, the future sequence of states and the discount [10, 35].

Dopamine-dependent RL models of addiction

Consistent with the neuropsychological view of addiction as habitual dominance in the decision making process, and inspired by reduced devaluation of drugs and pharmacological effects of drugs on dopamine [11, 76–78], several computational models have been proposed that focus on the MF process [35–37, 61].

The seminal work of Redish [61] 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. The resulting simulated behaviour reproduced several key features of addiction. In other words, this formulation hypothesized that drug values are always underestimated by the learning system, leading to unbounded growth of the value of drug related actions. This mechanism predicts that for drug related stimuli no blocking phenomena [81] should appear. However, blocking has been shown to appear both with natural and drug rewards [37, 82] falsifying this key prediction of the model.

A solution for this problem was first proposed by Dezfouli et al. [37]. Their study showed that if a different form of value was considered, averaged rather than discounted, then it would be possible to obtain similar results to those presented by Redish [61], while keeping bounded values for drugs and preserving the blocking phenomenon. Nonetheless, these studies [37, 61] do not account for the devaluation of drugs when followed by aversive effects (so called 'punishments') observed in the early stages of experiencing drugs [36, 83, 84].

Keramati and Gutkin [36] proposed an architectural explanation for addiction in

terms of hierarchical RL which allows one to account for blocking as well as for the devaluation of drugs in early drug-experience stages. Similarly to Redish's model [61], drug induced positive prediction error is present but it is repeated at several levels of the decision hierarchy and amplified when going from high to low levels of the hierarchy, which are classically associated with habits [17]. The authors propose that this hierarchical amplification would explain what is considered an inconsistency between cognitive and low-level evaluation of drug-related behaviours in addict individuals. It is suggested that, at the high (abstract) level, aversive effects associated with drug consumption overcome the drug induced positive prediction error bias, thus producing a negative abstract value. Conversely, at the low level, the opposite phenomenon is assumed to take place, as the amplification of drug induced positive prediction error overcomes the punishments, resulting in an overall positive low-level value. This hypothesis is also consistent with the general theory of habit dominance, where drug intake affects mostly low level decision components.

It has been suggested that Keramati and Gutkin's model [36], despite being conceived for a hierarchy of MF layers, could be extended to a hybrid architecture with MB, thus explaining non habitual drug oriented behaviours. However, it must be noted that this generalisation is unlikely because MB and MF have strong structural and generalisation difference. As already mentioned, the role played by dopamine as a value prediction error in MB planning is unclear. This model cannot generalise the overevaluation of drug-seeking behaviours in presence of drug related cues to situations which lack such cues. This is due to the interactions between high and low level components in the decision making. 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 addiction-like 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.

Finally, the theoretical proof of convergence to optimal behaviour of the standard MF models [69] may have intuitively suggested that a MF malfunctioning that impedes or delays convergence, caused by drugs [37, 61] or more complex architectural issues [36], would be necessarily part of a computational explanation for addiction. Nonetheless, convergence to optimal behaviour is supposed to be achieved after *infinite exploration of the environment* [69], a theoretical condition of limited interest when modelling biological phenomena. It will be shown that extremely long time for convergence and suboptimal addiction-like behaviours can have other causes than MF malfunctioning and dominance without calling into play MB impairments, in particular environment complexity.

Other RL based theories of addiction, such as the work of Simon and Daw [35], considered explicitly the duality aspects presenting both the MB and MF components. The starting point for such works is termed Dyna, one of the earliest dual models in AI [10, 72, 85, 86]. Dyna architecture assumes a shared value representation that would be updated by both the model-based and the model-free process. The model they propose [35] overcomes the issue that value prediction error encoded in midbrain dopaminergic signals does not seem to be involved in MB reward representation learning, which eliminates one of the possible direct vulnerabilities to drug of the MB process. In fact, such an architecture allows transferring drug-related vulnerabilities of

the MF dopamine system to the MB process. However, current data 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 [10].

Bounded cognitive resource of the MB components

A key issue of most dual models is that they assume an ideal, but expensive, model-based process [70, 87]. 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 [26, 27, 30, 88–91]. Thus, it is possible that these assumptions are hindering our ability to model biologically plausible healthy and dysfunctional (as e.g. in addiction [3, 35]) behaviours.

Even in the absence of any malfunction induced by drug consumption, e.g. drug reward over-estimation [3, 7], the following limits affect a MB system:

1. MB has computational limits and relies on costly computations. In fact, computing on the fly an optimal solution for realistic environments is out of reach of any computational system, including the brain [26, 27, 90]. Thus, the evaluation of long term consequences (value) of drug taking may not be trivial. The same computational limits may result in an extreme under-evaluation of non-drug related conditions, making irrelevant any negative long term effects of drug consumption. 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. MB performance can vary as a function of both environmental and internal conditions, such as the as the availability of computational resources. With limited computational resources, agents have to rely on approximated algorithms and each class of such algorithms has different shortcomings. As an example, search based algorithms [92–94] may be inefficient when action results are highly stochastic, as the evaluation of each course of action would require numerous samples. Therefore, MB assessment of long term effects of drug consumption may largely depend on the complexity of the environmental situation and the availability of computational resources (which might be limited for instance by stress or other cognitive loads);

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. Exploration is central for the AI RL literature especially in terms of the exploration-exploitation trade-off [19–21, 95–105]. Intuitively, this trade-off regards the decision between selecting either the action that maximises the total future expected reward based on current knowledge (i.e. the action with maximum value) or the one that minimises uncertainty thus possibly improving future decisions. Two main different objectives are proposed for this trade-off: achieving an efficient policy with a minimum number of samples (PAC [97, 99, 102]) versus minimising the overall cost of learning efficient policies (regret [101, 104, 105]). RL research in AI has also provided several measures of environment complexity (e.g. number of states, diameter [101] and Bellman rank [102]). The mechanisms that regulate the exploration/exploration trade-off in humans are not well understood. Most of the algorithms to control the exploration-exploitation trade-off proposed in AI and ML literature [97, 98, 100, 101] have only limited practical or modelling purpose

due to their high computational complexity. Thus only heuristic methods are available [56, 95, 96]. 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.

Materials and Methods

A Sketch of the Proposed Bounded Dual Model and Abstract Addiction-Enabling Environment

Agent Description. The model proposed in this paper makes minimal assumptions about the interaction between the MB and MF components. The action-values computed by the two process are combined with a fixed balance. Differently from Dyna [10, 72, 85, 86] there is no direct interaction between the MB and MF component during the computation of the action values. However, the two processes observe the same sequence of actions, states and rewards, which implies an indirect form of coupling mediated by the environment and behaviour.

The MB component has limited computational resources which are kept constant along a simulation in all the experiments.

Computational resources and the balance between MB vs MF represent some of important dimensions that determine individual learning and behaviours. They are used to define a spectrum of "simulated phenotypes" to study how they can impact on the risks and levels of addiction.

The agent implements a standard stochastic *persistent exploration* strategy [106] with low computational costs. This choice 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, thus it continues to perform exploratory actions with a non-zero frequency.

Environment Description. We hypothesise that addictive behaviours, characterised by *persistent and repeated* choices of a limited set of suboptimal actions, are induced by an environment structure where high immediate reward follows drug consumption, whereas limited punishment (negative effects of drug consumptions) is both unavoidable and distributed across many states following the drug selection. The abstract environment used in the experiment (Fig 2) 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.

Formal hypothesis

With the above theoretical background and methodological assumptions, we will focus on the following points:

1. Both MB and MF components can be affected by high exploration costs in complex environments. The exploration process under these conditions is intrinsically stochastic, resulting in significant variations in temporal extension. Stochastic behavioural trajectories can be divided in at least two classes with non zero probability, where the first converges relatively fast to the optimal behaviour, which we termed 'healthy', and the second diverges in its initial stochastic exploration segment, termed 'addicted'. In particular, initial exploration of the addicted class is characterised by: (i) longer duration in comparison with 'healthy' trajectories; (ii) repetition of few suboptimal actions. Interestingly, the addicted

class does not always result in addictive behaviour in the long term. We show *temporally extended transient exploration phases* which still eventually converge to optimal behaviours, without a drug induced alteration of the decision mechanisms;

2. Both MF and MB components contribute to addiction-like behaviours . A dysfunctional MB component plays an important role in shaping behaviours even in absence of drug related stimuli. This assumption is motivated by the unlikelyhood that generalizations of drug-related sub-optimal behaviours in drug-unrelated parts of the environment can be regulated by the MF omponent (e.g. as in relapse [22, 107]).

3. If addictive behaviours (re-)appear as a result of a change in the level of cognitive resources or a change of a part of the environment containing only 'natural' goals (RL optimal) then the MB component can play an important role in addiction even in absence of drug induced malfunctions. In fact, the MF component cannot explain the abrupt occurrence of sub-optimal behaviours in part of the environment (drug-related) as a consequence of failures in another part (drug-unrelated), which characterise several addiction and relapse conditions [22, 107];

4. Due to the above considerations, both MB and MF components of a dual decision system have limited reliability, specially in regard to exploration. Dominance of the goal oriented or of the habitual decision systems over the other may increase the probability of failure and addiction. Their *balance* could improve robustness to addiction because the two systems would probably not fail simultaneously as not all their vulnerabilities have common origins. A balanced MB & MF system may be more reliable to different maladaptive behaviours and malfunctions (e.g. lesions, see [17]) in addition to being more efficient.

Model Details

Fig 1. Dual Learning Agent

The behaviour of the agent used in the simulations (Fig 1) is controlled by a hybrid (or dual) model [66] that aims to maximise value (eq. 1) or expected cumulative reward, by simultaneously learning (MF) and computing (MB) an optimal action strategy, or policy π.

The MF component of the agent has been implemented using the Q-Learning algorithm in tabular form [69]. Q-Learning updates $Q^{MF}(s_t, a_t)$ value estimations after each action a_t execution using the formulas 2 - 3.

The MB component learns the statistics of the environment by explicitly keeping count of the transitions. The MB component is also *bounded* [30, 88] in terms of the amount of computational resources it employes to estimate the best course of action. To compute value estimation (Q^{MB}), this component applies at each step the Bellman Equation (4) 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 [108], with stochastic selection of the states update (see supplementary materials for details). The performance of the algorithm varies smoothly with the number of updates even if the environment presents several loops and sparse rewards.

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

Fig 2. 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 drug consumptions, 14 states associated with drug consumption 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 drug consumption and then enter the segment of states associated with drug consumption. 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 drug consumption 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 state. The overall average punishments received in the after-effects area make transiting to the drug consumption 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). Neutral states area is 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 Drug 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 drug area from state 8 and receives a reward of 10. Action results in drug areas are less predictable and result in randomly switch to other nearby states in the drug area. Only in state 15 the action a_w has higher chance of leaving the drug area. Leaving the drug area has a cost of -4 while every other transition inside the area costs -1.2. For an exact environment description see Table 2.

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:

$$Q^{MX}(s,a) = \beta Q^{MB}(s,a) + (1-\beta)Q^{MF}(s,a) \tag{5}$$

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 environment (Fig 2) has three types of states: healthy reward, neutral and drug. The agent deals with environment variations meant to simulate two conditions: drug condition and drug exposure (Fig 2.A, details in the supplementary material). Under the conditions of both first and second drug exposure ($d_{tpy} = 1000$ steps) the agent is rewarded by accessing both the healthy reward state (moderate reward, $R_g = 1$), and the drug segment (high reward, $R_d = 10$). Conversely, under the initial pre-drug condition ($d_{init} = 50$ steps), the agent does not receive any reward or punishment by entering the drug area (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}.

Table 1. Number of updates necessary to Prioritized Sweeping to find the optimal policy for each phase

Phase	Number of Updates
Init	4,712
Drug	5,005

Table 2. Environment Model Parameters

Name	Description	Value
N_T	Number of states	22
N_G	Number Goal States	1
N_D	Number Drug States	15
N_n	Number Neutral States	6
N_a	Number of actions	9
S_0	Starting state	4
R_p	Punishment end of drug consumption	-4
R_c	Punishment in drug area	-1.2
R_{dd}	Reward at drug consumption	10
R_g	Reward when entering goal state	1
d_{init}	Duration initial phase	50
d_{drug1}	Duration drug 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

Results

Fig 3. Evolution of healthy reward , rewarding drug and other states visit frequency with learning. The data are separated by MB/MF factor β and by addicted/healthy condition. Addicted agents visit drug (state 8, Fig 2) more often than healthy reward (state 1) by definition. The behaviours of the addicted agents are highly different from those of the healthy agents even with identical parametrisations, which is a form of behavioural bifurcation. Average trajectories over a population of 300 (healthy+addicted) subjects for each β value (3600 total simulated subjects). Smooth window size 75 steps. Addicted agents reach a maximum level of visits to healthy reward state before drug is introduced and then gradually stop visiting the healthy reward state (also due to smoothing effects). Conversely, healthy subject reach in both condition an average of 0.25 visit to the healthy reward state and have limited drug visit (0.05). Healthy agents with extreme values of β (0 or 1) are slower or reach lower healthy target visit frequency.

The exposure of simulated subjects to the environment described in figure 2 results in qualitatively different, almost opposite behaviours even with identical parametrisations (the β MB/MF balance) (see Figs 3-4). The behaviour trajectories may bifurcate between:

Fig 4. Evolution of healthy target/drug ratio for addicted and healthy subjects. The data are separated by MB/MF factor β and by addicted/healthy condition. Addicted agents visit drug state 8 more often than the healthy target state by definition. Average trajectories over a population of 300 (healthy+addicted) subjects for each β value (3600 total simulated subjects). Smooth window size 75 steps. Addicted agents reach a maximum level of visits to healthy reward state before drug is introduced (also due to smoothing effects) and then gradually stop visiting the healthy reward state. Conversely, healthy subjects achieve a ratio between drug 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.

healthy high (saturated) visit frequency to the healthy target and almost no visit to drug;

addicted almost no visits to healthy target and high visit frequency to drug.

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* only appear to belong to a stable condition reached after less than 300 steps that is actually only transient. Ultimately simulated addicted agents' behavioural trajectories will converge to optimal healthy behaviour. This was shown in longer simulations where convergence happens around 100K steps (Fig 5a), more than 300 times longer than the time necessary for the healthy agents to converge.

(a) Long runs with logarithmic time scale

(b) Ratio of addicted subjects varying β

Fig 5. a. Long runs with logarithmic time scale. Results show behaviour expressed by purely MF agents ($\beta = 0$). Average on 100 runs divided between those that 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 drug or other rewards. After 200K steps most of the addicted subjects go back to healthy, whereas addiction requires few tens of steps to initiate. **b. Ratio of addicted subjects with different balances between MF and MB components.** Total population of 300 simulated subjects. Addicted agents are those that during the observation period, 1000 time steps, acquire drug more often than healthy reward. Healthy subjects show the inverse behaviour. Intermediate values of β showed a lower ratio of addicted agents.

Simulated subjects that employ MB and MF components in a balanced manner show a higher resilience to addiction (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 5b). This is likely a result of higher robustness induced by having two parallel decision systems.

Figures 3-4 also show that healthy purely MB agents achieve a lower amount of healthy reward reaching frequency. This is due to using a bounded agent which cannot consistently compute the optimal action selection even with a correct model.

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. 4). 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 MB updates) results in 50% of the subjects addicted after 1K steps

(a) Initial Phase 1:1000 Steps **(b)** Complete 1:10,000 Steps

Fig 6. Number of addicted dependence on cognitive resources (MB 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 6a and 6b shows the effects of the exploration on all the parametrizations.

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 MB updates), 20% of the subjects show addiction at 1K steps falling to almost 0% after 10K steps to quit addiction (still 30 times longer than the time necessary for addiction to take place). 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. Furthermore, the selected amount of computational resources (500 MB updates) did not result in substantially worse performance (about 96%) than the optimal agent which stops at convergence (Table1) when the model was known. The change of performance, and the reduced addiction, results from both acquiring a correct model and consistently computing an efficient behaviour. 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, strong deliberative capacities, can help avoiding addiction, but a purely deliberative agent with high cognitive capabilities can still be susceptible of addiction due to bounded resources and exploration issues.

(a) Number of Addicted over 100 subjects

(b) State Visits Frequency Evolution (balanced subjects $\beta = 0.5$)

(c) State Visits Frequency Evolution (avg on identical parametrization subjects, both addicted and healthy)

(d) State Visits Frequency Evolution (extremes subjects $\beta \in \{0, 1\}$)

Fig 7. Effect of pre-training time (PTT, timesteps before drug introduction) and MB/MF balance factor β. Exposure to the environment before the introduction of drugs 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. This is confirmed by the higher focus on the healthy reward for high PTT shown Fig 7c.

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 choices and the early discovery of the rewarding states shapes the initial values estimation and biases future exploration.

Early acquisition of a correct model of parts of the environment that comprises the healthy rewards and a policy to achieve them, should 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 drug for 50, 100, 200, 500 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 reliable to addiction. This result may

provide the theoretical basis for current and novel prevention strategies.

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 drug 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.

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

Computational and animal models of addiction are often characterised by minimal environment complexity where the above described features are missing. This approach has led to important results, allowing to investigate the role of habitual and compulsive behaviour in addiction and its resistance to devaluation after the early stages [36, 83, 84, 109]. Nonetheless, addicted individuals in natural, non-controlled, environments face challenges and complex scenarios where the adverse consequences of drug consumptions are stochastic and distributed over long periods of time, reducing the chances of an individual to learn and model them comprehensively.

The simulations reported in the present study show that addiction-like drug-seeking behaviours can emerge and persist for a long time independently of failures of the learning mechanisms and as a result of the complexity of the environment.

Depending on less than *few hundreds* initial (stochastic) exploration steps, the simulated behaviour of the agents bifurcates. Eventually, two possible behaviours emerge, as each of the agents consistently and repeatedly either selects the healthy rewards in the environment or performs a maladaptive selection of the suboptimal drug-like outcomes, which may last up to *hundreds of thousands* of steps. This bifurcation phenomenon can be found with non-zero probability for any balance of control between MB and MF (β). The bifurcation is not due to a disruption of the classical TD-MF learning mechanism, which, differently from other works in the field [36, 37, 61], is intact in the simulated agents.

Crucially, these behaviours are the result of specific features characterising the environment. In particular, such determinant features are: (i) an easy access condition with a high reward, followed by (ii) difficult to estimate adverse consequences, such as delayed, stochastic or weak and prolonged negative rewards. Importantly, this

environmental pattern is not limited to the consumption of substances of abuse. 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 drug seeking actions. In fact, the agent, being 'attracted' by the high initial reward of drug, may neglect the exploration of competing and more convenient strategies and remain stuck in a disruptive behavioural 'local optimum'. On one hand, the MF component has substantial difficulties in propagating the 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. Thus, MF requires extensive exploration. On the other hand, the MB component is resource bounded and may fail in evaluating long term action effects when punishment spread across multiple states (feature ii) even after extensive training.

The presented results depend on the structure and complexity features of the proposed environment, which abstractly reflects addicted individuals' ecological conditions. It would have been possible to (i) define random environments, (ii) tune them to assume different values of complexity defined according to one of the several available metrics (as those defined in [101, 102]). Finally, (iii) expose dual RL agents that would show a spectrum of sub-optimal behaviours during learning. However, the set of environments with a given complexity is likely infinite and in each of those environments diverse suboptimal policies can appear. It is unlikely that a randomly selected environment from such set would allow a suboptimal policy containing an action that could be defined as 'addictive', i.e. repeated multiple times and followed by diverse but punishing and unavoidable consequences. So our approach appears more fit to model addiction.

Our results highlights that the use of oversimplified environments in previous studies may have lead to neglect the limits of the MB components and to overemphasize the role in addiction of an overactive MF component. In such conditions, an MB component is unlikely to induce suboptimal behaviour as values are easy to estimate and update. The proposed richer environment facilitates the analysis of the possible dysfunctions of MB decision processes and associated disorders in planning or exploring. To our knowledge this is an under-investigated factor for addiction. We hypothesise that higher robustness of balanced MB and MF agents is an effect of the different vulnerabilities that the two components suffer and their complementary strengths when facing complex environments. In fact the MB can be more efficient in terms of exploration and the MF can be robust to the occasional failures of the bounded MB in computing action values in a complex environment.

The present study focuses on the view of addiction-like behaviours as general phenomenon induced by environmental features interacting with the stochastic process of exploration and learning and not on the specific architecture or malfunctioning of the subject's decision system. The latter approach would face the issue of limited current knowledge on the details of the MF and MB processes implemented in the brain, especially regarding exploration and approximation methods employed, as well as the effects of bounded cognitive resources on real life decisions. The simulation results reported in this study are particularly relevant because they use a robust and classical architecture. While other models [65, 71] simulate a modulation of the balance β, our results are consistent over several values of β comprised the extreme 0 and 1, thus the observed phenomena should take place for any mechanism that modulates β. It would

be interesting in the future to test the presented results with different algorithms and architectures (e.g. diverse hierarchical structures and habit formation principles [36, 59, 111]) which environmental features can cause addiction-like behaviours. This may also help defining the environment features that induce suboptimal explorative behaviours and better match them with biological data.

It is difficult to relate our results to current experimental data on MB-MF balance. The presented simulations shows bifurcation between healthy and addicted behaviours for values of β both bigger and smaller values than 0.5 without any destructive pharmacological effect on the MF learning mechanism. Some studies have reported that a reduced contribution of the MB component may be a risk factor for addiction [110]. This may indicate that the subject in those experimental conditions would have a phenotype with values of β smaller than 0.5, where a reduction of the MB component induces higher risks of addictions (Fig 5b). In those studies, the balance is usually measured in a task specific manner by analysing behavioural responses. Using neural signals to extract a β is also complex because it is usually measured through the neural responses corresponding to MF prediction error which has actually shown to reflect both MB and MF components [66]. This estimation of β, however, does not correspond to the direct contribution of MB and MF to current action selection but it is related to learning the MF value. Thus, comparing the value β used in our model, which refers to a structural factor of the proposed model balancing MF and MB in action selection, with these estimations is not trivial.

The agents employed in the present study employs a persistent exploration strategy which is standard for changing environments. In Bayesian reinforcement learning and in active inference frameworks [20, 41], exploration is related to the prior expectations on the environment and the maximum reward achievable. In such frameworks it is not necessary to directly define an exploration factor, but the prior on the environment also dictates not only how the learning processes should change with progressing experience but also exploration. Moreover, active inference proposes a different (variational) take on bounded MB computations by restricting the space of behaviours and beliefs of the agent. The hierarchy of these spaces redefines the difference between habit and goal-directed processes, in place of MB vs MF, in a novel and more computationally unified manner [112]. Thus, active inference represents a logical choice for extensions of the current work to integrate exploration and environment induced issues with other classical factors of addiction disorder, e.g impulsivity.

Conclusions

In conclusion, a few statistical studies focus on the effects that different sources of complexity (most prominently, social factors [113, 114] and stress [115, 116]) may have on addiction, however current computational modelling literature mostly disregards these aspects [31–33]. In this work we have proposed a step forward in the simulation of complex environments with features that putatively and abstractly resemble the actual ecology of addiction.

In the future, an important research direction is the development of environment models more realistic than in the past for a more comprehensive computational understanding of addiction as well as other psychiatric disorders. 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. Till now, computational psychiatry studies [117, 118] have mostly focused on simplistic environments and optimality convergence conditions. As shown by the above simulations (in particular Fig 5a) those conditions may not be of interest for some disorders as they cannot be reached into a

subject lifetime. Conversely, considering complex environment features and subjects learning trajectory may reveal important missing factors in the understanding of mental disorders.

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. 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.
10. Daw ND, Dayan P. The algorithmic anatomy of model-based evaluation. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1655):20130478.
11. 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.
12. Norman DA, Shallice T. Attention to action: Willed and automatic control of behavior. Cognitive neuroscience: A reader. 2000; p. 376–390.
13. Nadel L. Multiple memory systems: What and why, an update. Memory systems. 1994;1994:39–63.
14. 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.

15. Thorndike EL. Animal intelligence: an experimental study of the associative processes in animals. The Psychological Review: Monograph Supplements. 1898;2(4):i.

16. Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998;37(4):407–419.

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

18. Balleine BW, Dezfouli A, Ito M, Doya K. Hierarchical control of goal-directed action in the cortical–basal ganglia network. Current Opinion in Behavioral Sciences. 2015;5:1–7.

19. 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.

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

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

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

23. 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.

24. 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.

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

26. 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.

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

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

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

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

31. 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.

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

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

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

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

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

37. 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.

38. Houk JC, Adams JL, Barto AG. A model of how the basal ganglia generate and use neural signals that predict reinforcement. 1995; p. 249–270.

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

40. Glimcher PW, Dorris MC, Bayer HM. Physiological utility theory and the neuroeconomics of choice. Games and economic behavior. 2005;52:213–256. doi:10.1016/j.geb.2004.06.011.

41. Schwartenbeck P, et al. Optimal inference with suboptimal models: addiction and active Bayesian inference. Medical hypotheses. 2015;84(2):109–117.

42. 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`.

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

44. 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.

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

46. 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.

47. Leyton M, Vezina P. Dopamine ups and downs in vulnerability to addictions: a neurodevelopmental model. Trends in pharmacological sciences. 2014;35(6):268–276.

48. Everitt BJ, Robbins TW. From the ventral to the dorsal striatum: devolving views of their roles in drug addiction. Neurosci & Biobehav Rev. 2013;37(9):1946–1954.

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

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

51. 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.

52. 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.

53. 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.

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

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

56. 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.

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

58. Niv Y. Reinforcement learning in the brain. Journal of Mathematical Psychology. 2009;53(3):139–154.

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

60. Daw ND, Doya K. The computational neurobiology of learning and reward. Curr Opin Neurobiol. 2006;16(2):199–204. doi:10.1016/j.conb.2006.03.006.

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

62. Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. 1999;12(7-8):961–974.

63. Asadi K. Strengths, Weaknesses, and Combinations of Model-based and Model-free Reinforcement Learning. Department of Computing Science University of Alberta; 2015.

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

65. 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.

66. 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.

67. 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.

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

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

70. 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.

71. 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.

72. 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.

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

74. Balleine BW, O'doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2010;35(1):48–69.

75. Isoda M, Hikosaka O. Switching from automatic to controlled action by monkey medial frontal cortex. Nature neuroscience. 2007;10:240–248. doi:10.1038/nn1830.

76. 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.

77. 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.

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

79. Kable JW, Glimcher PW. The neurobiology of decision: consensus and controversy. Neuron. 2009;63(6):733–745.

80. Deserno L, Huys QJ, Boehme R, Buchert R, Heinze HJ, Grace AA, et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proceedings of the National Academy of Sciences. 2015;112(5):1595–1600.

81. 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.

82. 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;.

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

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

85. 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.

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

87. 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;.

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

89. 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.

90. 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.

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

92. Kearns M, Mansour Y, Ng AY. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. Machine Learning. 2002;49(2):193–208.

93. Bonet B, Geffner H. Labeled RTDP: Improving the convergence of real-time dynamic programming. In: Proc. 13th Int. Conf. on Automated Planning and Scheduling; 2003. p. 12–31.

94. Kocsis L, Szepesvári C. Bandit based monte-carlo planning. Springer; 2006. p. 282–293.

95. 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.

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

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

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

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

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

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

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

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

104. 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.

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

106. Singh S, Kearns M, Litman DJ, Walker MA, et al. Empirical evaluation of a reinforcement learning spoken dialogue system. In: AAAI/IAAI; 2000. p. 645–651.

107. Deb P, Gallo WT, Ayyagari P, Fletcher JM, Sindelar JL. The effect of job loss on overweight and drinking. Journal of health economics. 2011;30:317–327. doi:10.1016/j.jhealeco.2010.12.009.

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

109. Pelloux Y, Everitt BJ, Dickinson A. Compulsive drug seeking by rats under punishment: effects of drug taking history. Psychopharmacology. 2007;194(1):127–137.

110. 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.

111. Donnarumma F, Maisto D, Pezzulo G. Problem Solving as Probabilistic Inference with Subgoaling: Explaining Human Successes and Pitfalls in the Tower of Hanoi. PLOS Comput Biol. 2016;12(4):e1004864.

112. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O'Doherty J, Pezzulo G. Active inference and learning. Neuroscience and biobehavioral reviews. 2016;68:862–879. doi:10.1016/j.neubiorev.2016.06.022.

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

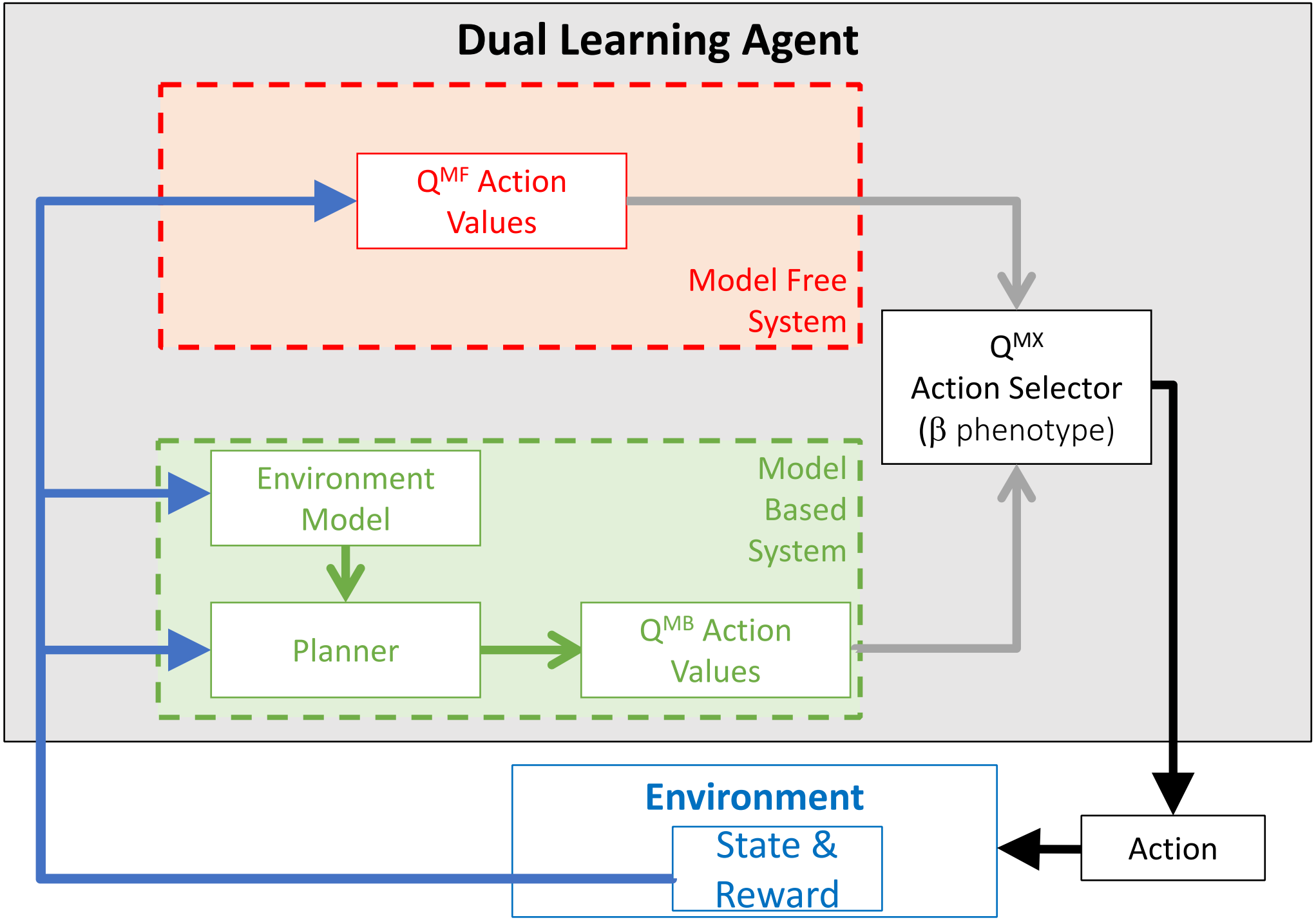
114. Havassy BE, Hall SM, Wasserman DA. Social support and relapse: Commonalities among alcoholics, opiate users, and cigarette smokers. Addictive behaviors. 1991;16(5):235–246.

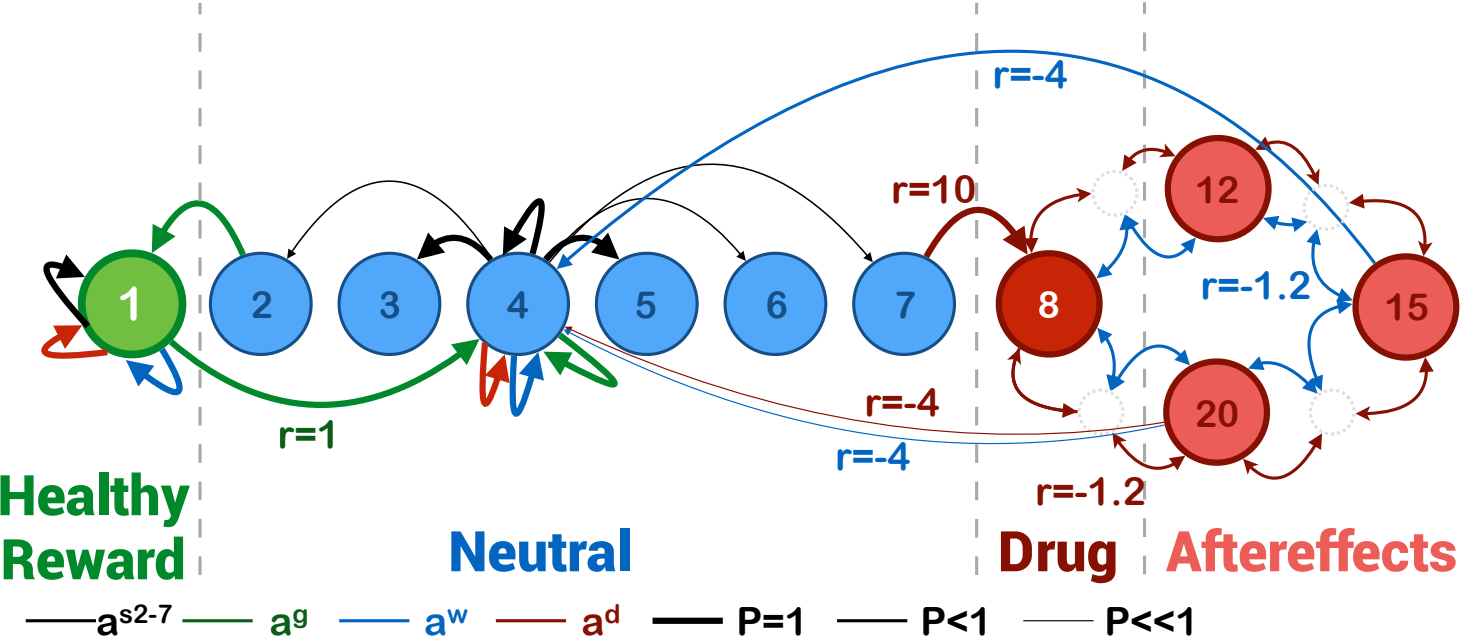
115. Sinha R, Shaham Y, Heilig M. Translational and reverse translational research on the role of stress in drug craving and relapse. Psychopharmacology. 2011;218(1):69.

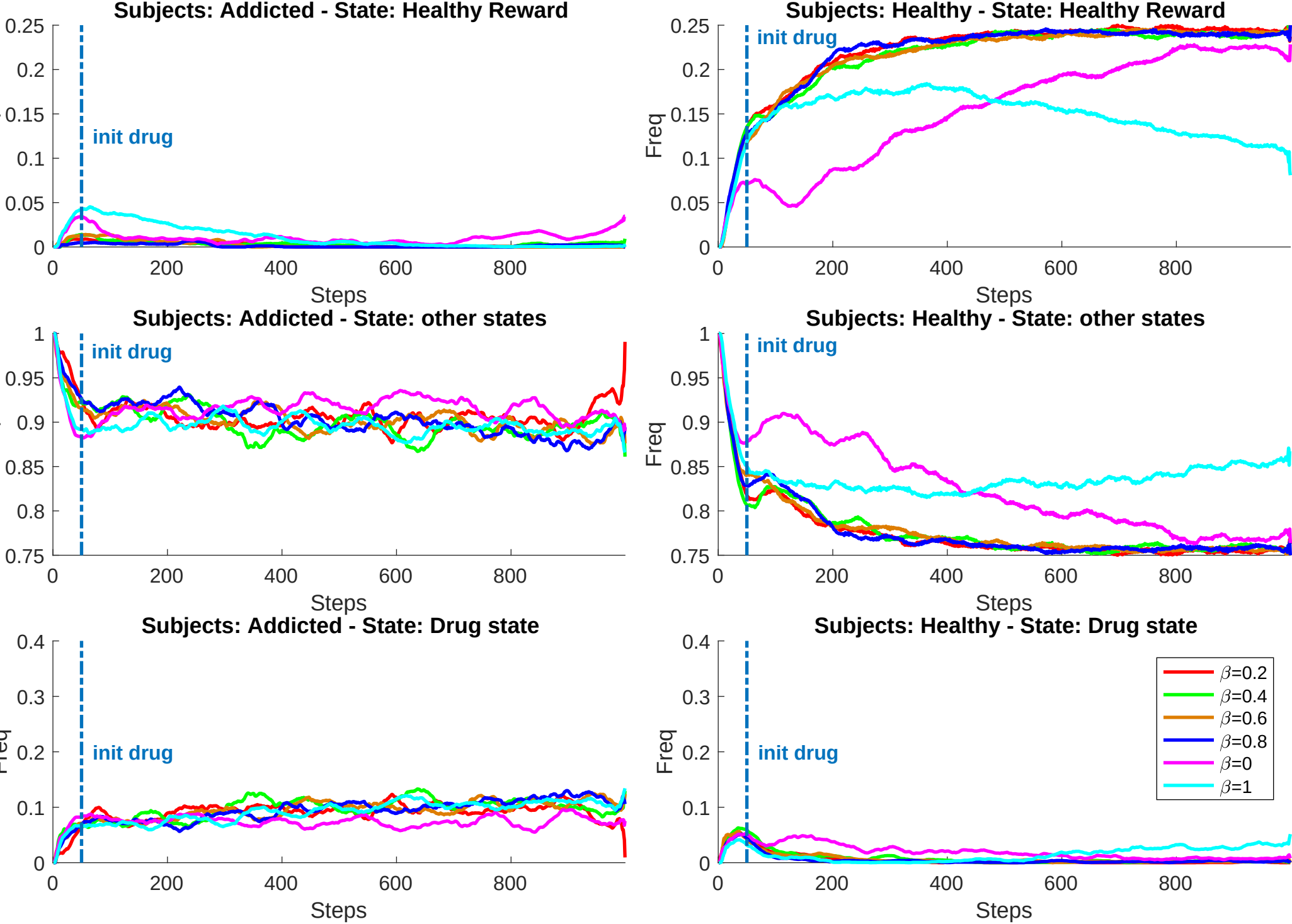
116. Mantsch JR, Baker DA, Funk D, Lê AD, Shaham Y. Stress-induced reinstatement of drug seeking: 20 years of progress. Neuropsychopharmacology. 2016;41(1):335–356.

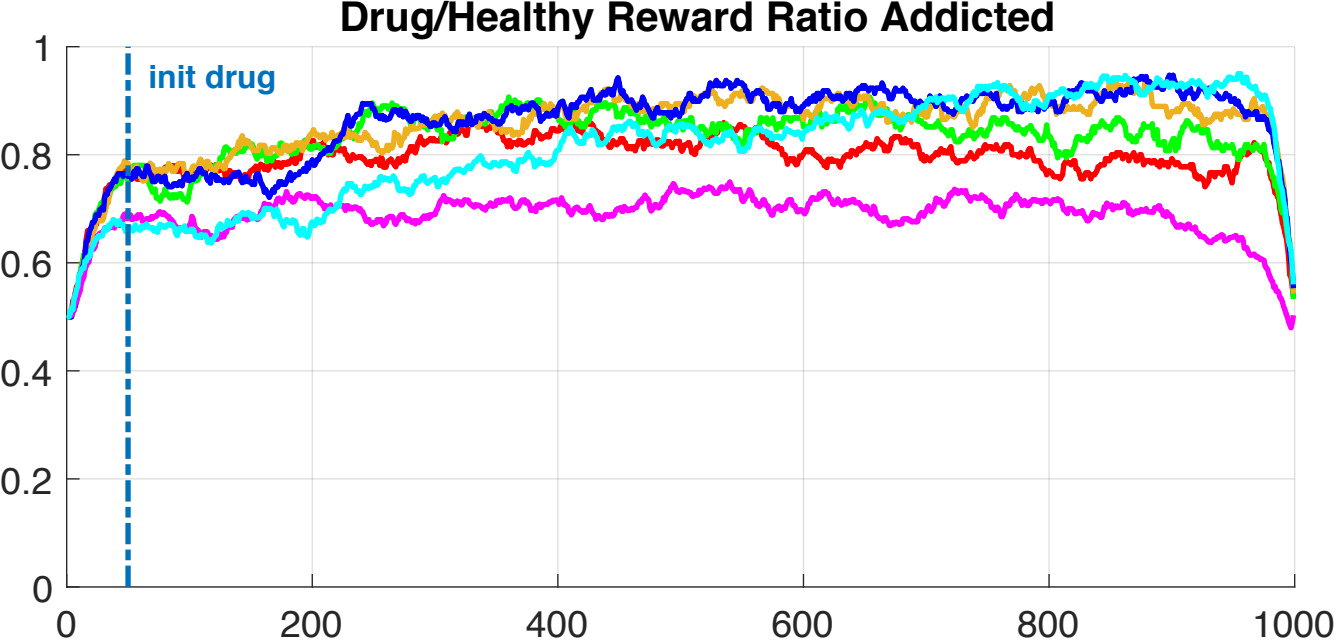
117. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends in Cognitive Sciences. 2012;.

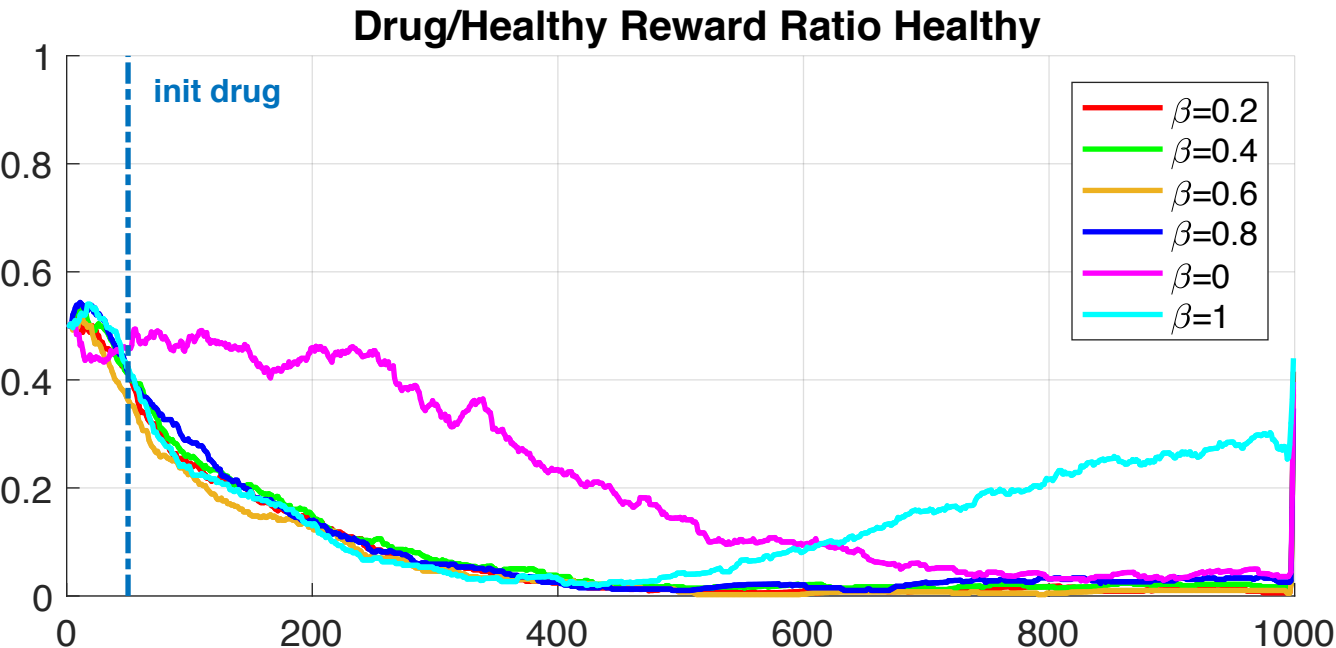
118. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014;84(3):638–654.

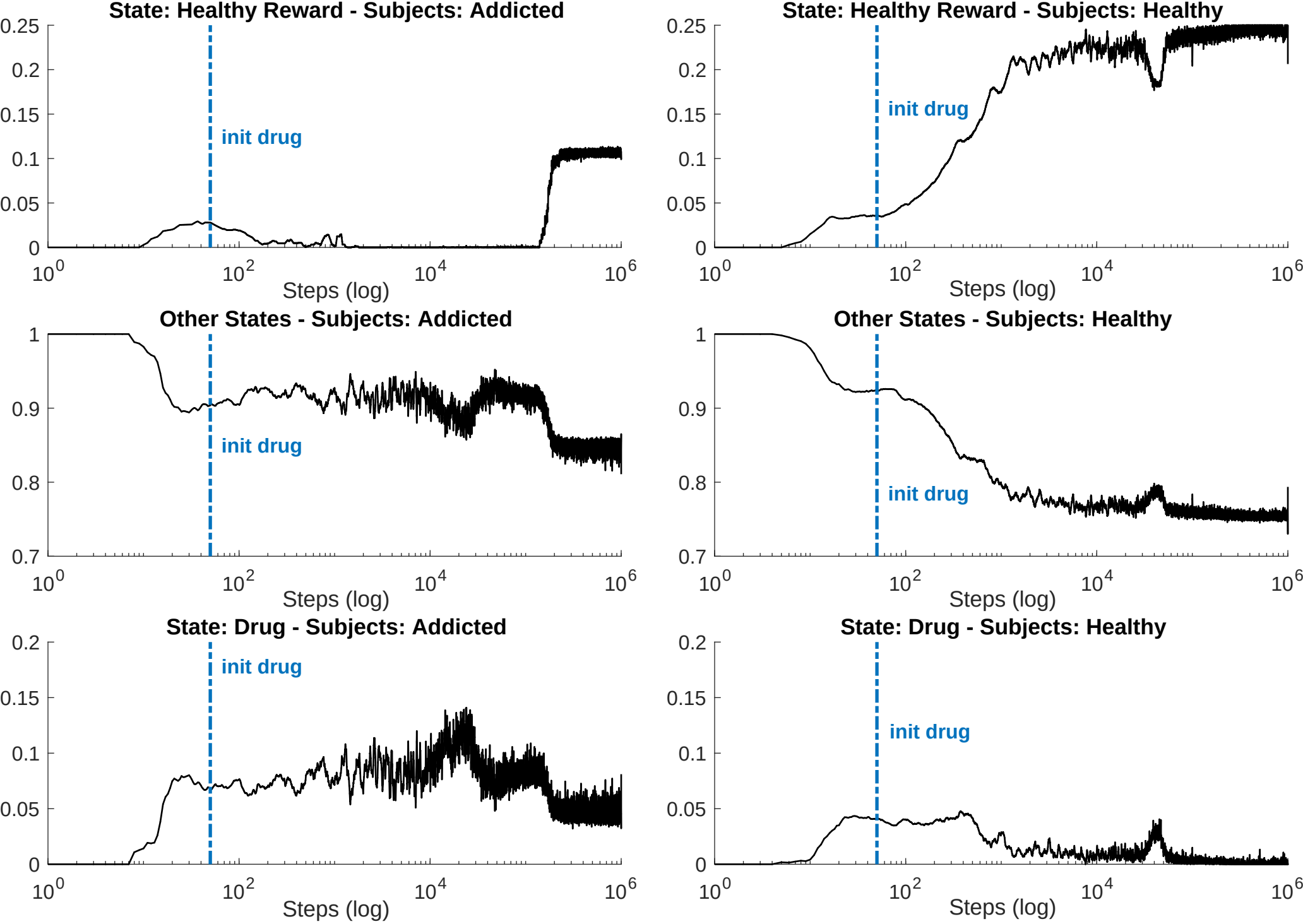

Dual Learning Agent
Q^{MF} Action Values
Model Free System
Q^{MX} Action Selector (β phenotype)
Environment Model
Model Based System
Planner
Q^{MB} Action Values
Environment
State & Reward
Action

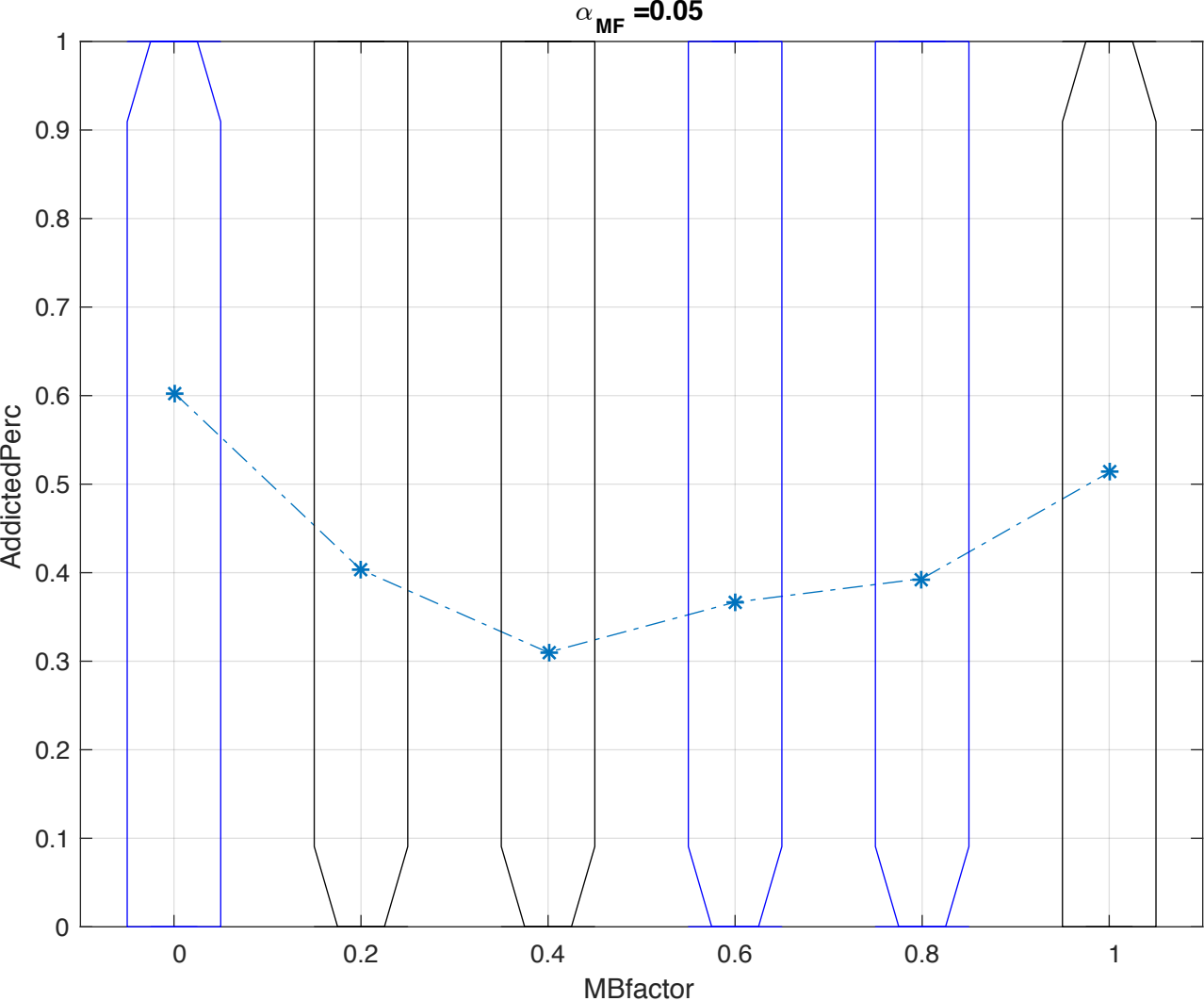

r=-4
r=10
1
2
3
4
5
6
7
8
12
15
20
r=-1.2
r=-4
r=1
r=-4
r=-1.2
Healthy Reward
Neutral
Drug
Aftereffects
a^{s2-7}
a^{g}
a^{w}
a^{d}
P=1
P<1
P<<1

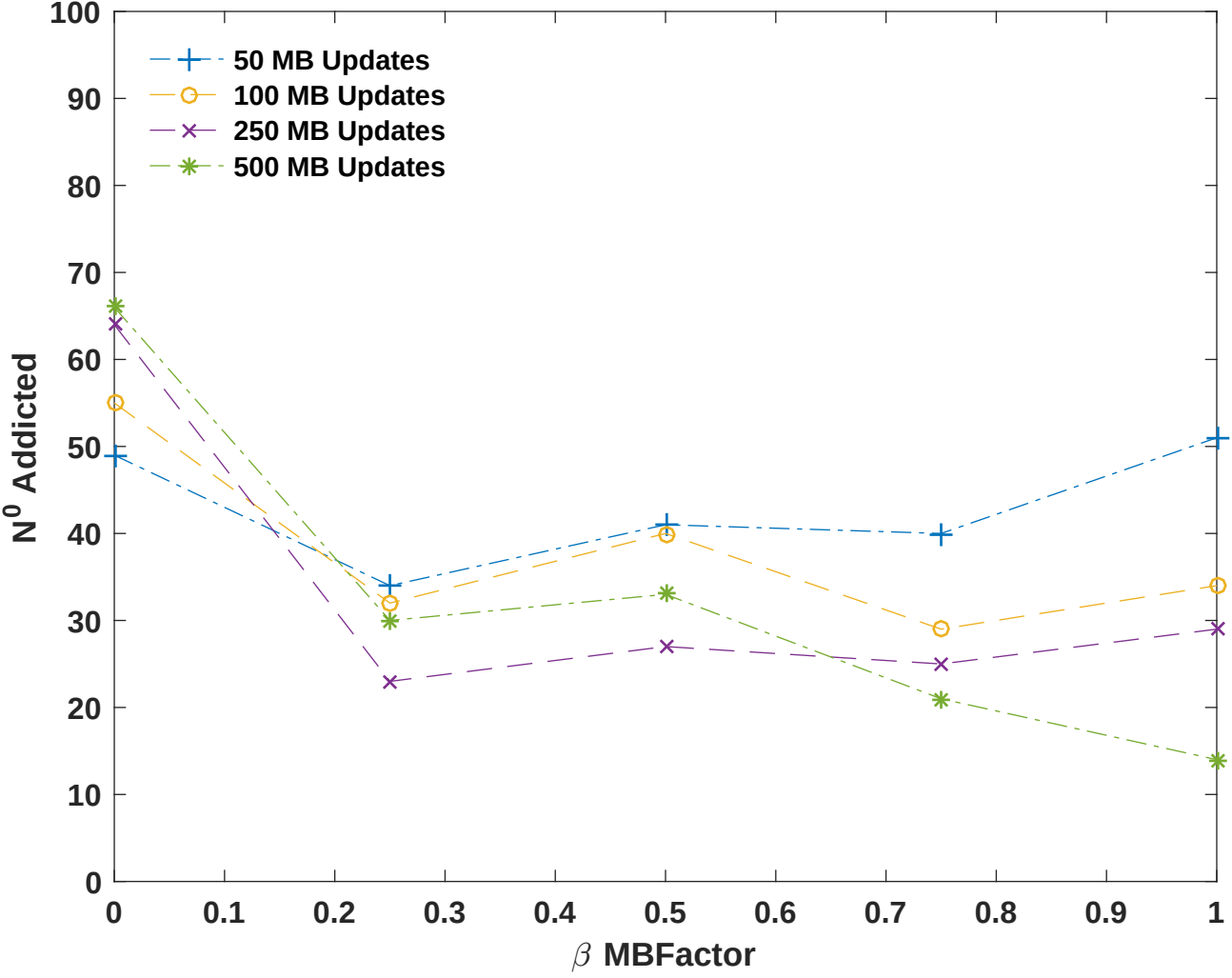

Subjects: Addicted - State: Healthy Reward
Subjects: Healthy - State: Healthy Reward
Subjects: Addicted - State: other states
Subjects: Healthy - State: other states
Subjects: Addicted - State: Drug state
Subjects: Healthy - State: Drug state
init drug
Freq
Steps
β=0.2
β=0.4
β=0.6
β=0.8
β=0
β=1

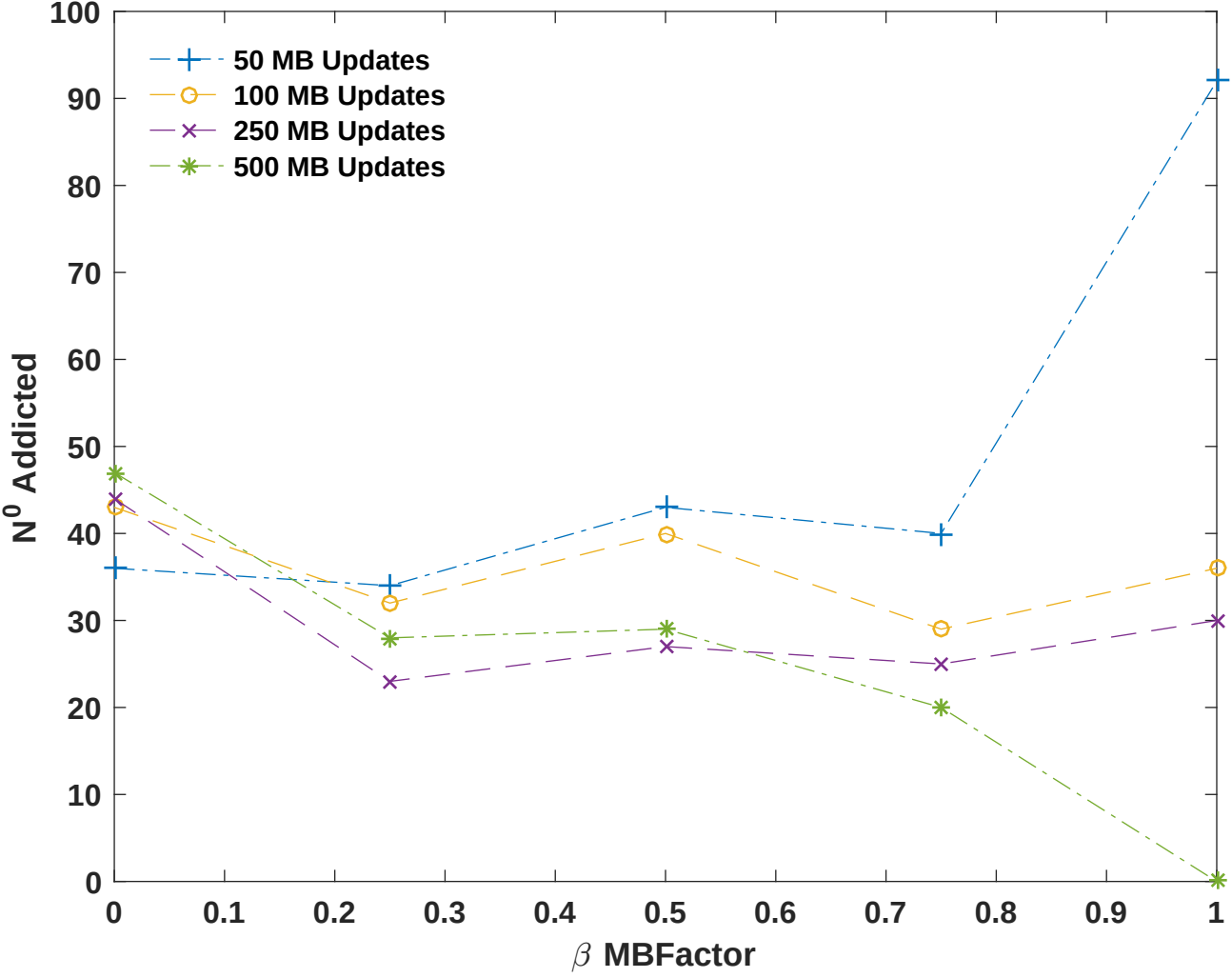

Drug/Healthy Reward Ratio Addicted
init drug
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000

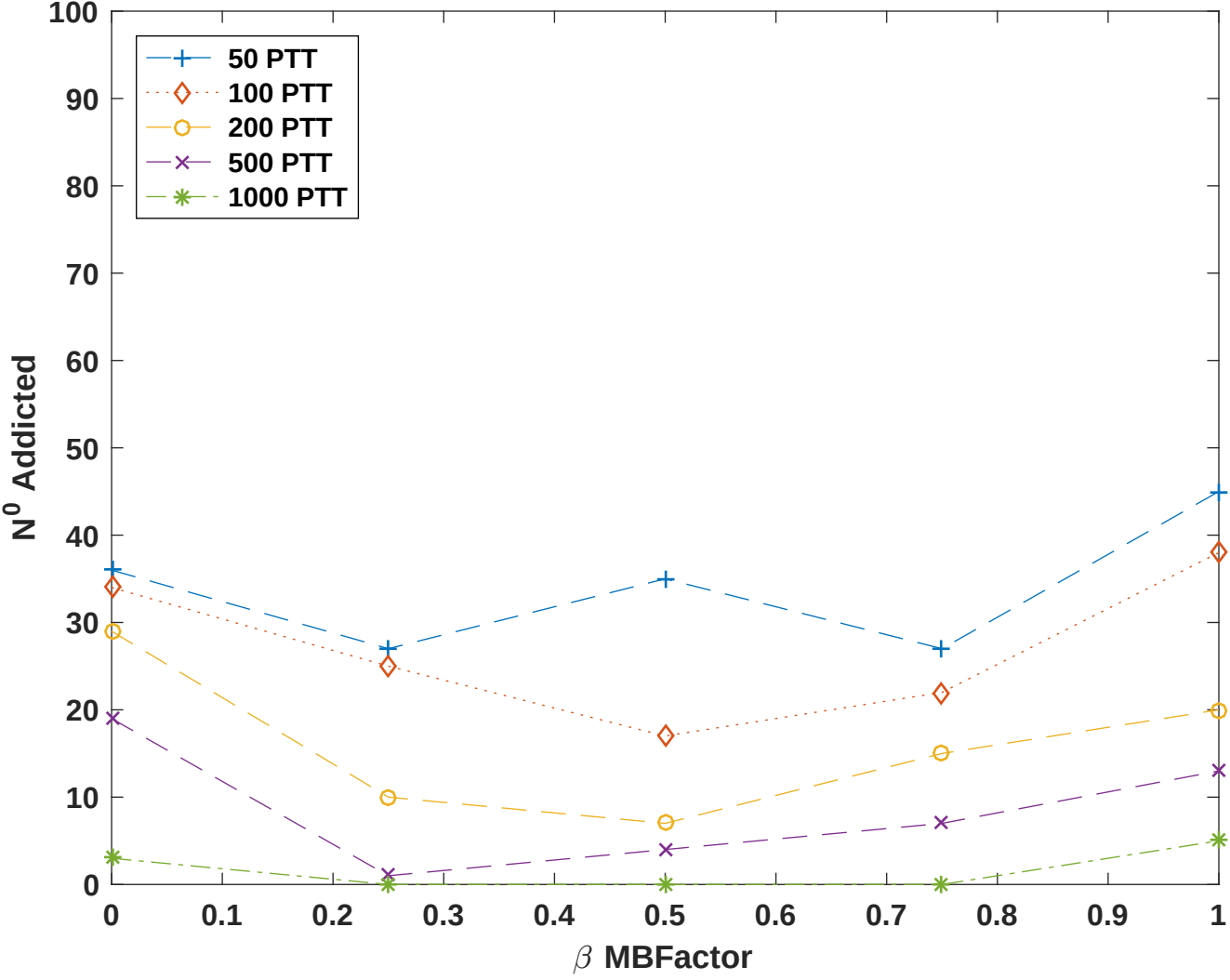

Drug/Healthy Reward Ratio Healthy
init drug
β=0.2
β=0.4
β=0.6
β=0.8
β=0
β=1
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000

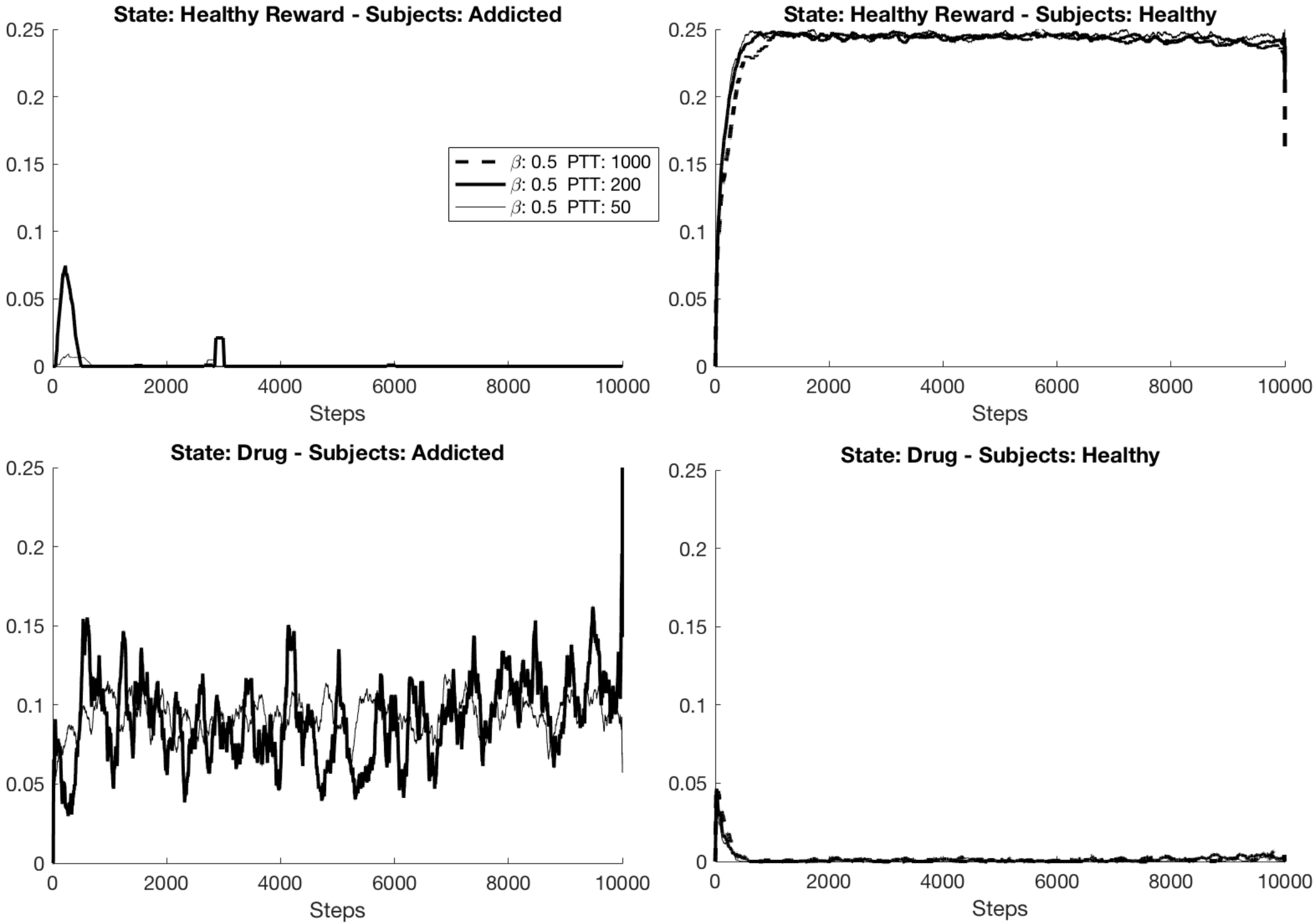

State: Healthy Reward - Subjects: Addicted
State: Healthy Reward - Subjects: Healthy
Other States - Subjects: Addicted
Other States - Subjects: Healthy
State: Drug - Subjects: Addicted
State: Drug - Subjects: Healthy
init drug
Steps (log)
0.25
0.2
0.15
0.1
0.05
0
1
0.9
0.8
0.7
10^0
10^2
10^4
10^6

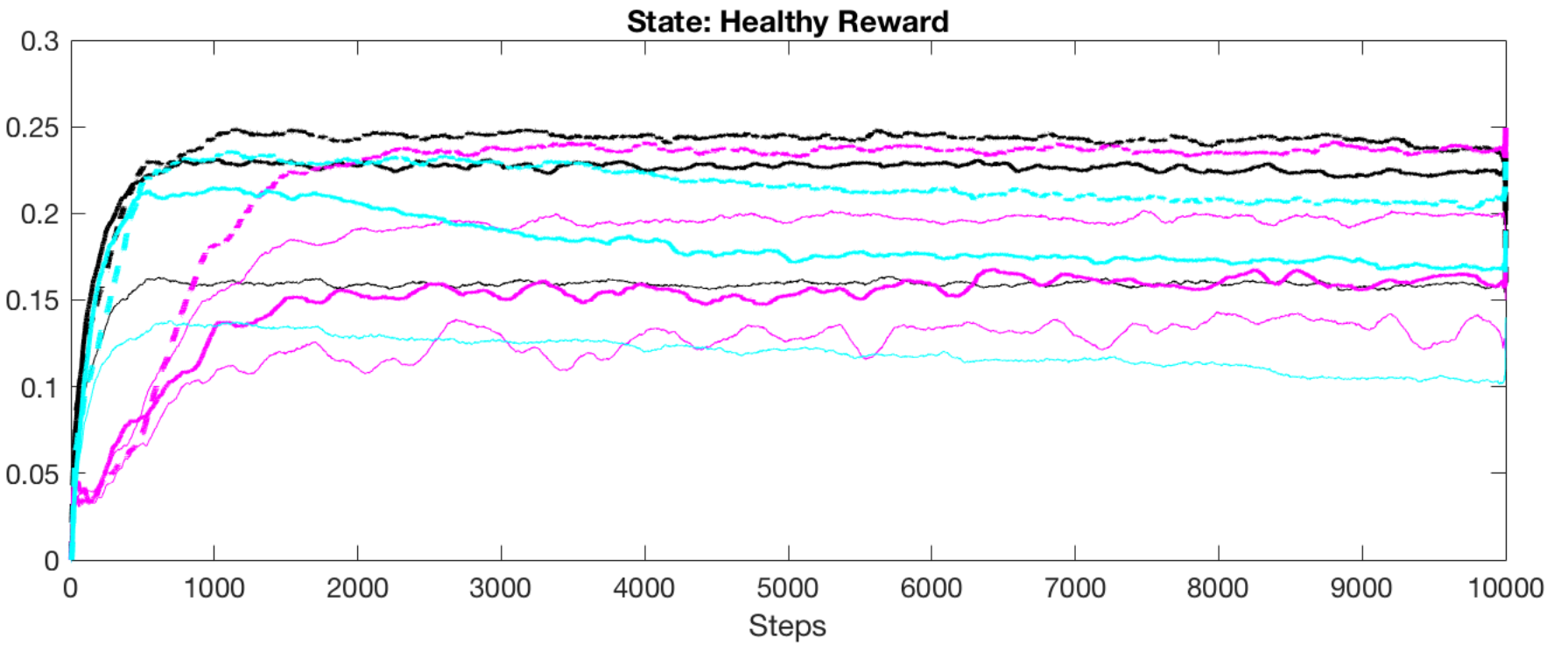

α_{MF} =0.05
AddictedPerc
MBfactor
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1

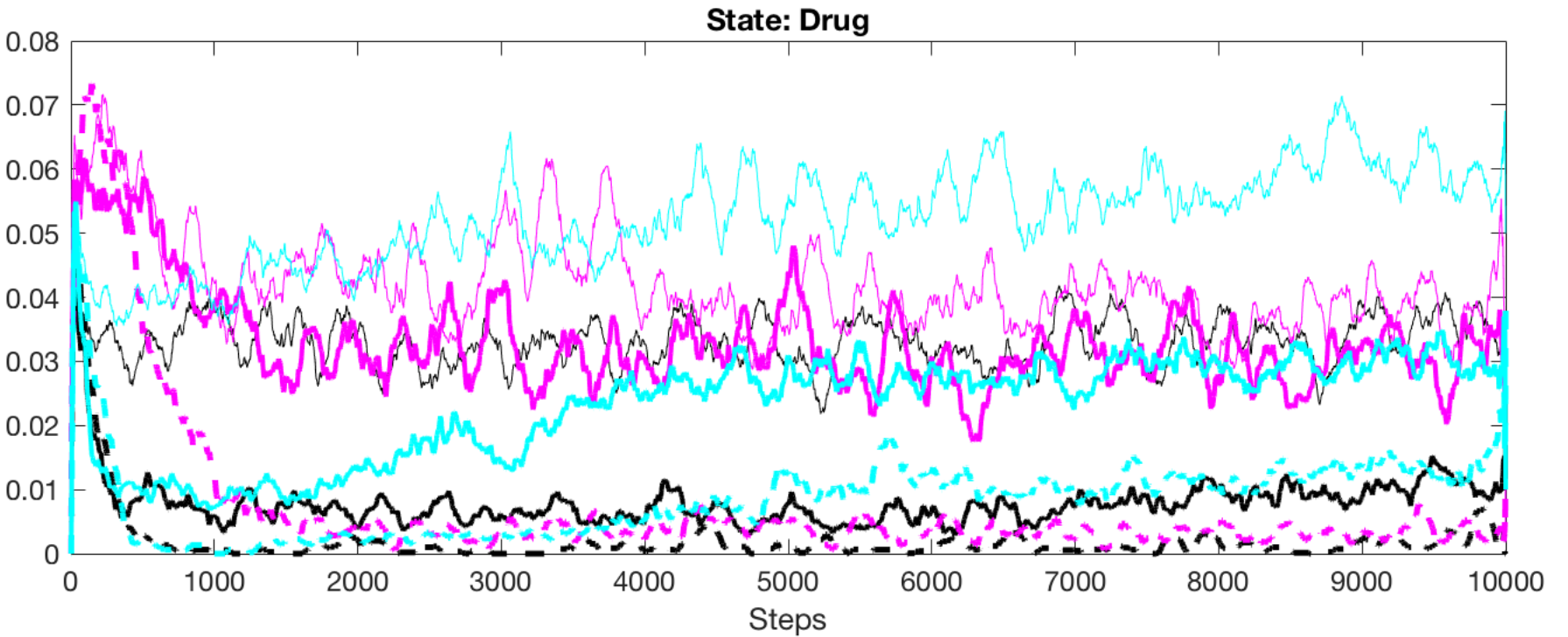

50 MB Updates
100 MB Updates
250 MB Updates
500 MB Updates
N0 Addicted
β MBFactor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

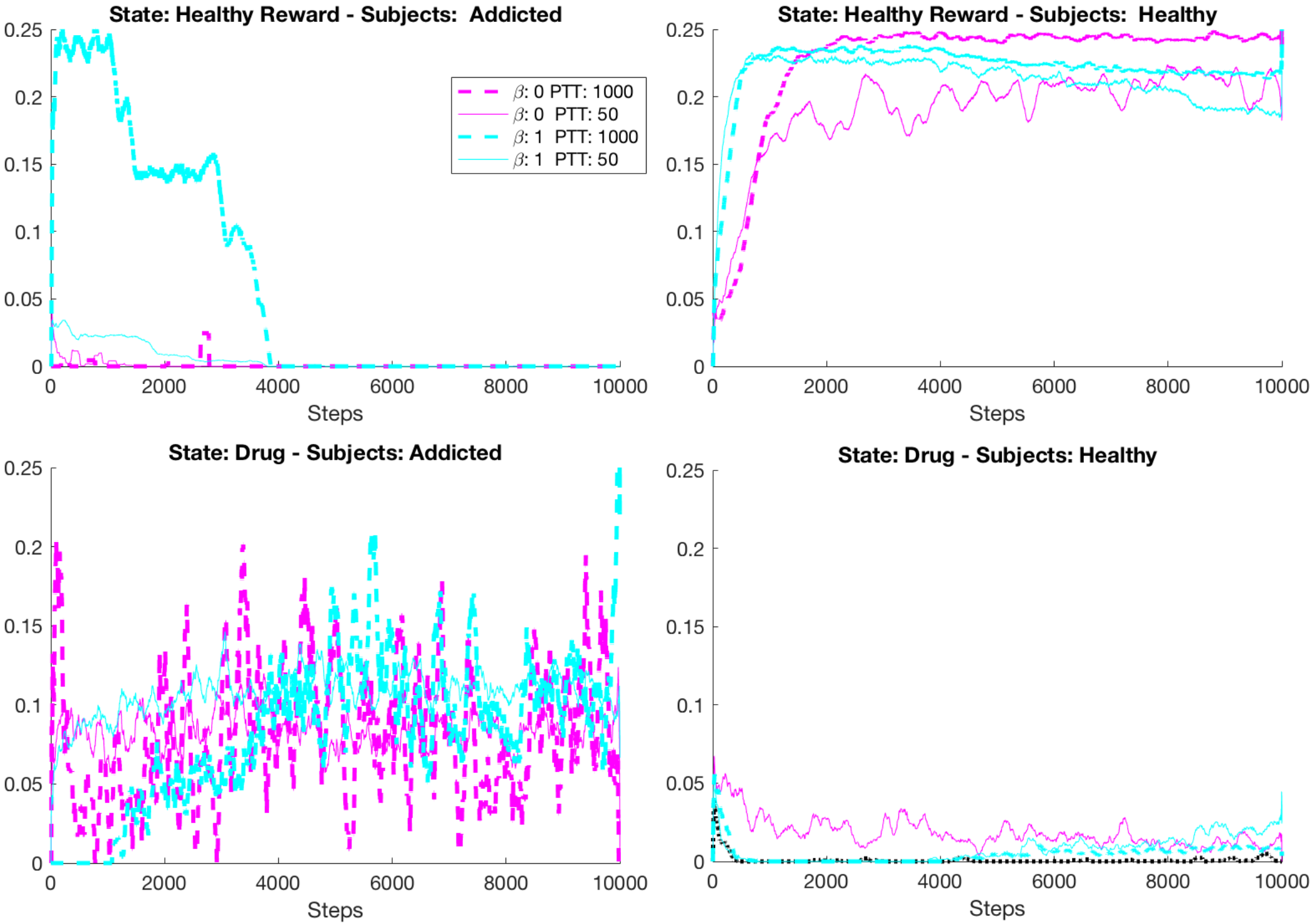

50 MB Updates
100 MB Updates
250 MB Updates
500 MB Updates
N0 Addicted
β MBFactor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

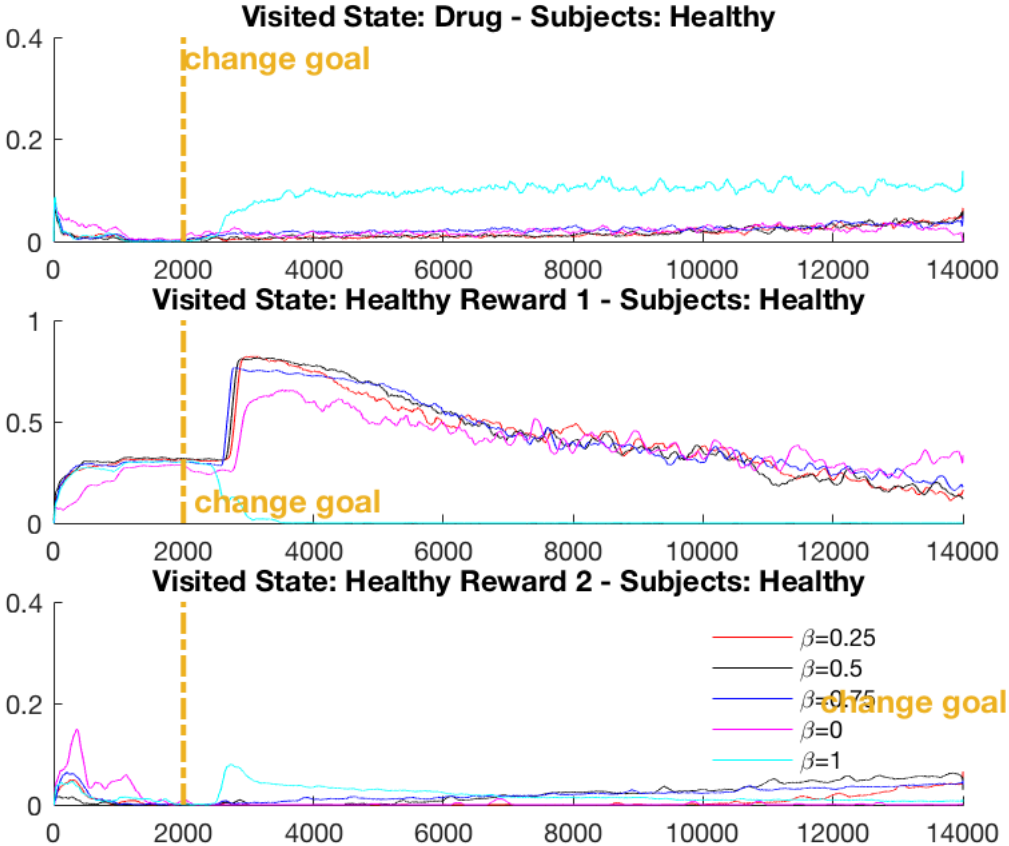

50 PTT
100 PTT
200 PTT
500 PTT
1000 PTT
N^0 Addicted
β MBFactor
0
10
20
30
40
50
60
70
80
90
100
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


State: Healthy Reward - Subjects: Addicted
State: Healthy Reward - Subjects: Healthy
State: Drug - Subjects: Addicted
State: Drug - Subjects: Healthy
β: 0.5 PTT: 1000
β: 0.5 PTT: 200
β: 0.5 PTT: 50
0.25
0.2
0.15
0.1
0.05
0
0
2000
4000
6000
8000
10000
Steps


State: Healthy Reward
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Steps


State: Drug
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Steps


State: Healthy Reward - Subjects: Addicted
State: Healthy Reward - Subjects: Healthy
State: Drug - Subjects: Addicted
State: Drug - Subjects: Healthy
β: 0 PTT: 1000
β: 0 PTT: 50
β: 1 PTT: 1000
β: 1 PTT: 50
Steps
0
0.05
0.1
0.15
0.2
0.25
2000
4000
6000
8000
10000


Visited State: Drug - Subjects: Healthy
change goal
0.4
0.2
0
0
2000
4000
6000
8000
10000
12000
14000
Visited State: Healthy Reward 1 - Subjects: Healthy
1
0.5
0
change goal
Visited State: Healthy Reward 2 - Subjects: Healthy
β=0.25
β=0.5
β=0.75
β=0
β=1
change goal