1   **Manuscript Title:** A Multilevel Computational Characterization of Endophenotypes in

2   Addiction

3

4   **Authors and Affiliations**: Vincenzo G. Fiore[1*], Dimitri Ognibene[2,3], Bryon Adinoff[4,5], Xiaosi

5   Gu[1*]

6   [1] School of Behavioral and Brain Sciences, University of Texas at Dallas, 800 W. Campbell

7   Road, Richardson, TX 75080-3021, USA

8   [2] Department of Computer Science and Electronic Engineering, University of Essex, Colchester,

9   CO4 3SQ, UK

10   [3] Department of Information and Communication Technologies, Universitat Pompeu Fabra,

11   08018, Barcelona, ES

12   [4] University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390,

13   USA

14   [5] VA North Texas Health Care System, 4500 S. Lancaster Rd., Dallas, TX 75216 USA

15

16

17   [*] Corresponding author:

18   Vincenzo G. Fiore, Vincenzo.g.fiore@gmail.com

19   Center for BrainHealth, School of Behavioral and Brain Sciences, UT Dallas

20   2200 West Mockingbird Lane, Dallas, TX 75235, USA. Tel: +1 214.905.3007

21

22

23

24   **Abstract**

25   Substance use disorders are characterized by a profound intersubject (phenotypic) variability in

26   the expression of addictive symptomatology and propensity to relapse following treatment.

27   However, laboratory investigations have primarily focused on common neural substrates in

28   addiction, and have not yet been able to identify mechanisms that can account for the

29   multifaceted phenotypic behaviors reported in literature. To investigate this knowledge gap

30   theoretically, here we simulated phenotypic variations in addiction symptomology and responses

31   to putative treatments, using both a neural model based on cortico-striatal circuit dynamics, and

32   an algorithmic model of reinforcement learning. These simulations rely on the widely accepted

33   assumption that both the ventral, model-based, goal-directed system and the dorsal, model-free,

34   habitual system are vulnerable to extra-physiologic dopamine reinforcements triggered by drug

35   consumption. We found that endophenotypic differences in the balance between the two circuit

36   or control systems resulted in an inverted U-shape in optimal choice behavior. Specifically,

37   greater unbalance led to a higher likelihood of developing addiction and more severe drug-taking

38   behaviors. Furthermore, endophenotypes with opposite asymmetrical biases among cortico-

39   striatal circuits expressed similar addiction behaviors, but responded differently to simulated

40   treatments, suggesting personalized treatment development could rely on endophenotypic rather

41   than phenotypic differentiations. We propose our simulated results, confirmed across neural and

42   algorithmic levels of analysis, inform on a fundamental and, to date, neglected quantitative

43   method to characterize clinical heterogeneity in addiction.

44

47   **Introduction.**

48   Substance use disorders are known to encompass a wide range of individual behavioral

49   differences (i.e. phenotypes) in addiction development, maintenance and severity of symptoms,

50   and treatment response (Everitt and Robbins, 2016). Previous investigations into the mechanisms

51   underlying this heterogeneity of behaviors have identified two fundamental alterations that are

52   correlated with a vulnerability in the development and severity of substance consumption

53   (Garrison and Potenza, 2014; Jupp and Dalley, 2014; Belin et al., 2016). These neural and

54   computational differentiations (i.e. endophenotypes) are respectively identified with a

55   dysregulation of D2 receptors in the striatum (Morgan et al., 2002; Nader and Czoty, 2005;

56   Dalley et al., 2007; Volkow et al., 2007; Flagel et al., 2010; Belcher et al., 2014; Flagel et al.,

57   2014; Gould et al., 2014) and of learning rates (Gutkin et al., 2006; Piray et al., 2010). However,

58   these endophenotypic differences are found across a wide spectrum of dissociable phenotypes, so

59   that the same neural or computational mechanism is used to account for separable behavioral

60   traits. For instance, striatal D2 dysregulation is found in individuals differing in terms of their

61   impulsivity (Dalley et al., 2007; Volkow et al., 2007), social dominance (Morgan et al., 2002;

62   Nader and Czoty, 2005; Gould et al., 2014), motor reactivity or preference for novelty (Flagel et

63   al., 2010; Flagel et al., 2014) or sensitivity to rewards (Belcher et al., 2014). Each of these

64   behavioral traits is separately correlated with development of addiction, but they do not

65   necessarily coexist in the same individuals (cf. novelty seeking and impulsivity: Ersche et al.,

66   2010; Molander et al., 2011; Belin and Deroche-Gamonet, 2012; Flagel et al., 2014). This

67   mismatch between few known endophenotypic differences and a wide variety of multifaceted,

68   dissociable, behavioral phenotypes suggests there are yet unknown neural and computational

69   mechanisms that are responsible, alone or in interaction, for the reported behavioral

3

70   differentiations. Finally, investigations into intersubject variability often emphasize the initial

71   stage of addiction development (but see e.g.: Belin et al., 2008; Economidou et al., 2009; Pelloux

72   et al., 2015). Yet, individual differences also emerge in response to treatments, resulting in

73   diverse relapse patterns, among individuals showing similar severity of symptoms. These

74   differences have not been so far addressed in neural or computational models.

75

76   To overcome these shortcomings, we propose a theoretical investigation into the interaction

77   among ventral/dorsal cortico-striatal circuits and the associated behavioral control modalities.

78   We have developed two models to simulate neural dynamics and algorithmic (or normative)

79   choice selections in a multiple-choice task involving drug and non-drug rewards. Consistently

80   with previous models, we assume addictive substances hijack the healthy reward prediction error

81   signal (Schultz et al., 1997) by triggering extra-physiologic dopamine bursts (Nestler and

82   Aghajanian, 1997; Koob and Volkow, 2016). These dopamine activities signal the presence of an

83   aberrant unexpected reward, leading to the repetition of drug-related actions and escalation of

84   consumption (Redish et al., 2008; Dayan, 2009). In our neural model, this process of

85   reinforcement learning (RL, Sutton and Barto, 1998) is mediated by extra-physiologic changes in

86   cortico-striatal connectivity weights (Hyman et al., 2006; Haber, 2008). These changes in turn

87   aberrantly affect circuit gain and the stability of both ventral and dorsal cortico-striatal circuits,

88   disrupting their respective role in encoding goal-directed behaviors (Balleine, 2005; Balleine and

89   O'Doherty, 2010; Gruber and McDonald, 2012) and habitual responses (Yin et al., 2004;

90   Balleine and O'Doherty, 2010). A similar effect is assumed for our algorithmic model, where

91   over-evaluation of drugs and related RL affect the two control modalities, termed *model-based*

92   and   *model-free*,   that   loosely   correspond   to   ventral/goal-oriented   and   dorsal/habitual

93   implementations (Dolan and Dayan, 2013; Voon et al., 2017). As a result, and consistently with

94   previous formulations of RL models of addiction (Redish et al., 2008; Piray et al., 2010; Gillan et

95   al., 2016), both the planned evaluation of known action-outcome contingencies, represented in an

96   *internal model* of the world, and the reactive immediate motor responses are biased towards

97   drug-related selections.

98

99   Based on these assumptions, our models show that phenotypic differentiation in addiction

100   development and treatment response can emerge as a function of a variable regulating the

101   interaction between ventral and dorsal circuits or model-based and model-free control modalities.

102   Our simulated results offer a proof of concept that this interaction is a candidate independent

103   neural and computational mechanism underlying addiction vulnerability, putatively

104   characterizing three different endophenotypes differing in the likelihood to develop addiction,

105   severity of symptoms and treatment response. We suggest this neurocomputational mechanism

106   could interact with previously described ones to generate the variety of dissociable behavioral

107   traits reported in literature as associated with addiction vulnerabilities.

108

109   **Materials and Methods.**

110   In brief, we present two complementary models simulating endophenotypic differences and their

111   effects on addiction development and treatment response. We tested our simulated agents under

112   four conditions, and in environments granting free access to a substance of addiction, as usually

113   implemented in laboratory studies. In particular, we compared our simulated phenotypic

114   variability with the results described in a recent study investigating individual differences in rats

115   self-administrating the stimulants cocaine or a designer drug, a dopamine- and mixed dopamine-

116   norepinephrine reuptake inhibitor, respectively (Gannon et al., 2017). We selected this study

117   because it highlights how different drugs, dosages and tasks result in different ranges of

118   phenotypic differentiation. For instance, an initial acquisition phase, over a 10-day period, shows

119   compulsive behavior developed in up to 75% rats self-administering cocaine and 87.5% of those

120   exposed to the designer drug. Furthermore, under a condition of fixed ratio (=5) schedule, the

121   study showed self-administration varied significantly among subjects. A subset of rat population,

122   termed high responders, self-administered cocaine up to 60% more times in comparison with a

123   different subset, termed low responders, depending on dosage (cf. figure 3 in: Gannon et al.,

124   2017). Thus, our focus was on perturbing the balance between the dorsal/model-free and the

125   ventral/model-based systems, to compare our simulated behavioral differentiations in drug

126   consumption with the data reported in the chosen laboratory study.

127

128   The two models comprised a neural mass model that has been validated and described in the

129   context of choice behavior and dopaminergic modulation (Baldassarre et al., 2013; Fiore et al.,

130   2014; Fiore et al., 2016) and a normative or algorithmic model based upon standard RL schemes.

131   In the neural model, addiction and treatment response were modeled through DA-dependent

132   associative plasticity in both ventral and dorsal circuits. In the RL model, aberrant learning was

133   modeled using a duplex of model-based and model-free schemes that competed for control over

134   action selection. The model-based scheme entails learning a model of the environment (in the

135   form of probability transition matrices among states) that was used to compute value functions

136   under the Bellman optimality principle. The equivalent model-free scheme used prediction error

137   based learning to directly acquire the value of state action pairs. Both neural and RL models

138   were tested under four successive phases; 1) before exposure to drug (termed *pre-drug*); 2)

6

139 learning of addictive behavior (termed *addiction*); 3) therapeutic interventions (termed

140 *treatment*) that revert the learning of the previous phase. These are stylized interventions that

141 simulate treatments prevalently affecting either goal-orented/model-based or habitual/model-free

142 control systems. In the first case, the treatment is assumed to modify only the internal model of

143 the environment and related selection of action-outcome contingencies performed in the ventral

144 circuit. In the second case, the treatment represents a condition in which the model of the world

145 of the agent remains mainly unaltered, but the acquired drug-related stimulus-response

146 associations are disrupted, thus preventing the agent from exhibiting habitual responses (cf. Doll

147 et al., 2009). Finally, 4) under the last condition we reinstated access to the simulated drug

148 following each treatment (termed *relapse*). The unique aspect of this complementary modeling

149 approach is that neural and algorithmic models provide construct validation of each other, as

150 process and implementation theories (i.e., synaptic and dynamical mechanisms) complement the

151 normative principles formalized in the RL model.

152

153 ### *1. Neural field model.*

154 *Basic model architecture and parameterization:* In cortico-striatal circuits, the signal processed

155 in the cortex is conveyed towards its respective area of the striatum, processed in basal ganglia

156 and finally relayed to the same cortical area where it originated, via thalamus (Haber, 2003;

157 Draganski et al., 2008; Jahanshahi et al., 2015). Thus, despite diverging in terms of the

158 information processed –e.g. sensorimotor or rewards and outcomes– these circuits are

159 characterized by similar computational dynamics (Obeso et al., 2014). Temporal responses in

160 recurrent neural networks co-occur with state transitions or input transformations that are often

161 described in terms of energy landscapes (**Fig. 1A-C**). If multiple inputs or initial states generate

162   transitions towards the same final state, this is termed *attractor state* (Amit, 1989). In recurrent

163   networks such as cortico-striatal circuits, learning processes modulate the circuit gain, thereby

164   affecting the strength of the attractor states and the overall stability of the system (Fiore et al.,

165   2015).

166

167   We simulated the temporal responses in cortico-striatal circuits in a neural model developed in

168   Matlab (illustrative representation of the neural architecture is represented in **Fig. 1D**). This

169   neural model simulates mean-field activity (Deco et al., 2008) within multiple channels of both

170   dorsal and ventral cortico-striatal loops. A continuous-time differential equation (equation 1)

171   simulates changes over time (τ_g) of the average action potential (u_j) of a pool of neurons, and a

172   positive transfer function (equation 2) converts this action potential in the final activation of the

173   pool (y_j). Finally, the plasticity of the connections (w_{ij}) between cortex and striatum is

174   characterized by DA-dependent Hebbian learning, corrected with a constant threshold (*th*) as

175   defined in equation 3. The resulting rule strengthens the connections among all active nodes in

176   the cortex and those active in the striatum, and weakens the connections among nodes showing

177   opposite activation status.

178

179    $\tau_g \dot{u}_j = -u_j + b_j + (\epsilon + \lambda d) \sum w_{ji} y_i$ (1)

180    $y_j = [\tanh(u_j - \theta)]^+$ (2)

181    $\Delta w_{ij} = \eta \left([y_i - th]^+ [y_j - th]^+ [d - th]^+ \right) - \zeta \left([th - y_i]^+ [th - y_j]^+ [d - th]^+ \right)$ (3)

182

183   The input $(\sum w_{ji} y_i)$, reaching each node in the neural network is modulated by two coefficients λ

184   and ϵ. These regulate the ratio between the signal affected by the presence of dopamine release d

8

185   and the amount of signal that is computed independent of dopamine release. For most units, the

186   values of the two coefficients are set to $\lambda = 0$ and $\epsilon = 1$, with the exception of the simulated

187   striatal units, where these parameters are set to $[\lambda = 1.4, \epsilon = 0.2]$ and $[\lambda = -0.5, \epsilon = 0.6]$, to

188   simulate the differential effect dopamine has, depending on the most prevalent receptor type

189   (λ>1 and λ<0 for D1 and D2 receptors, respectively). Due to the different effects the dopamine

190   receptors have on the activity of the simulated neurons, the drug-induced dopamine-dependent

191   Hebbian learning significantly affects D1-enriched units in the striatum, whilst having negligible

192   effects on D2-enriched units (Gerfen and Surmeier, 2011; Volkow and Morales, 2015).

193

194   *Simulating different addiction phenotypes and treatment effects*: Agents controlled by the neural

195   model are immersed in a simplified environment and can select among three arbitrary actions or

196   inactivity (cf. non stationary three armed bandit environment). The selection of the actions is

197   carried out in the circuit simulating the dorsal cortico-striatal activity and it is considered

198   completed if the neural activity of any of the units in the external layer of the simulated cortex

199   (cf. Fig.1) is maintained for at least 2 seconds. Ventral and dorsal circuits interact, both ways, via

200   cortico-cortical connectivity. Therefore, the activity in the simulated ventral circuit biases action

201   selection in the dorsal circuit and the selection of actions in the dorsal circuit biases the activity

202   in the ventral circuit. To test our hypothesis about the effect these reciprocal biases have on

203   choice behavior, we assumed cortico-cortical weights do not vary over time and we tested eleven

204   combinations for the parameters determining their weights, as w_{ji}=[0.02-0.2], [0.03-0.17], [0.05-

205   0.15], [0.07-0.13], or [0.1-0.1] (and symmetrical). This spectrum of weights describes the

206   strength of the biases between the two major circuits, thereby characterizing either a balanced

207   condition or a dominance of one of the two circuits. We report the effects in terms of behavioral

208   responses for these putative endophenotypes and test each of these with thirty noise seeds,

209   random inputs and under four conditions, to allow within phenotype comparisons. The first

210   condition, "pre-drug", represents an assessment of behavior before any drug or reward is

211   introduced, as the three available inputs randomly change their value to determine a non-

212   stationary order of preferences. Under the second condition, termed "addiction", one action is

213   associated with the administration of a simulated addictive substance, triggering DA phasic

214   responses and associated Hebbian learning in cortico-striatal connections of both ventral and

215   dorsal circuits. For the third condition, termed "treatment", we simulate the effects of deprivation

216   coupled with one of two hypothetical treatments targeting either the dorsal or the ventral cortico-

217   striatal circuits. The treatments are simulated reverting the learning process in either the dorsal or

218   the ventral cortico-striatal circuit. The dorsal treatment brings back the pre-drug configuration in

219   the dorsal circuit and keeps the configuration reached under the condition of addiction for the

220   ventral circuit. The ventral treatment is achieved with the opposite intervention. Finally, under

221   the fourth condition, termed "relapse", we reintroduce access to the addictive substance,

222   inducing relapse. Under this condition, relapse time is defined as the time required to reinstate

223   the configuration of cortico-striatal weights found at the end of the *addiction* condition.

224

225   *2.* **RL model.**

226   *Basic model architecture and parameterization*: In this model, we assume that the behavior of

227   the agent relies on a hybrid model (Daw et al., 2011) that learns and computes the value of

228   choices (actions, a_t) under each condition (state, s_t). Value is defined as a quantity that combines

229   short and long term expected rewards and negative outcomes when a specific strategy of action

230   is followed (policy, π). It is formally defined as:

231

232 $\qquad Q^{\pi}(s_t a_t) = r(s_t a_t) + E\left[\sum_i^{\infty} \gamma^i\, r\big(s_{t+i}, a_{t+i} = \pi(s_{t+i})\big)|s_t, a_t\right]$ (4)

233

234 In equation (4), $r(s, a)$ denotes the instantaneous reward received when action a is performed in

235 state s. γ is a discount factor, comprised between 0 and 1, which defines the trade-off between

236 immediate and long term rewards. The value of a state given the policy is defined as $V^{\pi}(s) =$

237 $\max_a Q^{\pi}(s, a)$. For each environment there is an optimal policy $\pi^*(s)$, which maximizes the

238 value $V^{\pi^*}(s)$ for every state (Sutton and Barto, 1998).

239

240 The environment can be completely characterized through the state transitions distributions

241 $p(s_{t+1} = s|s_t, a_t)$, and the expected rewards $E(r|s, a) = R(s, a)$. These two functions together

242 represent a model of the environment. Model-based behaviors compute $Q^{\pi}(s_t a_t)$ and the policy

243 relying on such functions, at each state, following the Bellman equation (Daw and Dayan, 2014):

244

245 $\qquad Q^{\pi^*}(s_t, a_t) = R(s_t, a_t) + \gamma \sum_s \left[p(s_{t+1} = s|s_t, a_t) \max_a Q^{\pi^*}(s, a) \right]$ (5)

246

247 The model-based component learns the transition distributions and the expected rewards during

248 the interaction with the environment. Thus, differently from other hybrid models (Daw et al.,

249 2005; Keramati et al., 2011; Pezzulo et al., 2013), the quality of Q value estimation at any given

250 moment depends on the experience the agent acquired up to that point in time. To compute value

251 estimation (Q_{MB}), this bounded (Gershman et al., 2015) component applies at each step the

252 Bellman equation (5) a limited number of times ($N_{PS} = 50$) to states sampled stochastically

253 following a heuristic for efficient state update selection. The algorithm is an early-interrupted

11

254   variation of the Prioritized Sweeping algorithm (Moore and Atkeson, 1993) with stochastic state

255   update selection. Crucially, our model-based component does not accumulate the variations of Q

256   values over time, and restarts the computation after each step (desJardins et al., 1999). This

257   choice is meant to instate a plausible bounded rationality for our model which can account for

258   the cognitive costs and ensuing limits of integrating old and new information about the

259   environment, whilst updating and extending a complex plan to navigate it. This implementation

260   was suitable for a bounded rational model-based component that shows controlled stochasticity

261   of deliberation performances in non-trivial environments. This choice allowed to test the effects

262   of the hypothesized endophenotypic differentiation in an environment characterized by higher

263   degree of complexity in comparison with both the one chosen for the neural model and those

264   described in the literature of RL models of addiction. In particular, we considered drug

265   consumption to be associated with complex after-effects that make it difficult to predict the

266   overall result of pursuing the related action course.

267

268   In comparison with other hybrid models such as Dyna and Dyna2 (Sutton, 1990; Silver et al.,

269   2016), the proposed architecture does not share Q values between model-based and model-free

270   components, nor it requires that the two processes share the same state representations. The two

271   components separately represent their Q values and integrate them in a later phase. This

272   decoupling is assumed to result in a more biologically plausible agent (Daw & Dayan 2014), and

273   it is essential for the simulations of two separate treatments, essential requirement to establish a

274   comparison with the behavior simulated with the neural model. In contrast with previous work

275   using a hybrid Dyna-like architecture and Prioritized Sweeping algorithm, where the sharing of

276   the Q-values explained the appearance of model based drug oriented behavior (Simon and Daw,

277　　2012), in our model this model based addiction emerges in independent model-free and model

278　　based components. Thus, addiction behavior results from the joint effect of high reward (i.e. the

279　　drug), a limited number of stochastically selected policy updates and limited knowledge of the

280　　environment.

281

282　　The model-free component has been implemented using the Q-Learning algorithm in tabular

283　　form (Watkins and Dayan, 1992). Q-learning updates initial state value estimations as follows:

284

285　　$Q_{MF}(s_t, a_t)_{new} = Q_{MF}(s_t, a_t)_{old} + \alpha \delta_t$ 　　　　　　　　　　　(6)

286　　$\delta_t = R(s_t, a_t) + \gamma \max_{a'}[Q_{MF}(s_{t+1}, a')] - Q_{MF}(s_t, a_t)$ 　　　　　(7)

287

288　　where α is a learning factor comprised between 0 and 1. Our hybrid model computes choice

289　　values in a fashion that balances model-free (MF in the equations) and model-based (MB in the

290　　equations) components depending on a parameter β. Six values (1, 0.8, 0.6, 0.4, 0.2, 0) are used

291　　for this parameter to simulate different endophenotypes, on a spectrum between purely model-

292　　based (β=1) and purely model-free (β=0) RL.

293

294　　To allow exploration, the action to execute is selected randomly 10% of the times. This

295　　exploration factor is kept constant to support adaptation to a changing environment (Singh et al.,

296　　2000) and to simulate the continuous update of knowledge necessary to cope with ecological

297　　environments. The remaining 90% of the times, actions are determined by maximizing $Q_{MX}(s,a)$

298　　in a strategy defined as ε-greedy (ε=.1). These values are produced by combining the values

299　　computed by the model-based and model-free components:

300

301 $$Q_{MX}(s,a) = \beta Q_{MB}(s,a) + (1-\beta)Q_{MF}(s,a) \tag{8}$$

302

303 The choice for a fixed balance between model-based and model-free requires minimal

304 assumptions on their interaction and has been used in recent reinforcement learning architectures

305 (Silver et al., 2016).

306

307 *Simulating different addiction phenotypes and treatment effects:* In comparison with the

308 simulations characterizing the neural model, for the RL model we could develop a more complex

309 environment, to highlight how our endophenotypic differentiations could also affect the

310 likelihood to develop addiction. This environment is characterized by a total of 20 states divided

311 into four different types (**Fig. 2**): (i) healthy rewards (i.e. normal rewards that are not directly

312 associated with drugs); (ii) neutral states (no reward or negative outcome); (iii) drug-related

313 states, which give a high reward but are followed by multiple (iv) drug aftereffects, characterized

314 by small negative outcomes. Similar to the neural model investigations, the agent deals with

315 environment variations meant to simulate four phases of addiction: initial pre-drug phase (f1);

316 addiction (i.e. the drug becomes accessible for the first time, f2); treatment (f3); relapse (i.e.

317 second drug exposures, f4). Under the initial pre-drug phase (d_{init}=50 steps), the agent does not

318 receive any reward or negative outcome by entering the drug-related and aftereffects area, but a

319 moderate reward is assigned (R_g=1) by accessing the healthy reward state. Under the conditions

320 of addiction and post-treatment addiction (d_{tpy}=1000 steps), the agent can also receive a high

321 reward, after accessing a drug-related state (R_d=10). The drug state always leads to a series of

322 randomized state transitions among the aftereffects states (R_a=-1.2) and simulates generic

323   negative consequences associated with addiction. The agent can occasionally leave this

324   aftereffect area of the environment (**Fig. 2**) to reach a neutral state, at the price of a further

325   negative outcome (R_a=-4). Under the treatment phase (d_{tpy}=1000 steps), the drug-related state

326   results in a negative outcome (Rdt=-1, see **Tables 1** and **2**, column f3), thus increasing the

327   chances the agent stops pursuing this state. To allow for a comparison with the results in the

328   neural model, we simulated a model-based and model-free treatment by manipulating the

329   learning factor of the non-treated control modality, decreasing it: α_{Ctpy}=0.01 * α. Under the

330   condition of relapse, we measure the simulated time required by the agents to reach at least 95%

331   of drug-related action preference as recorded under the addiction condition, after the drug is

332   introduced again in the environment. This threshold is used to measure the percentage of agents

333   relapsing, as well as the time required to complete the relapse, per endophenotype.

334

335   **Results.**

336   **Simulations from the neural field model**.

337   Under all conditions, the three stimuli randomly change every few seconds, putatively

338   representing a dynamic fluctuation of values associated with perceived cues in a non-static

339   environment. Optimal behavior requires the agents to rapidly adapt to these changes, transiently

340   triggering the motor response associated with the most valuable cue (cf. Fiore et al., 2016;

341   Hauser et al., 2016). Under pre-drug condition, dorsal and ventral circuits perform unbiased

342   selections, collaborating in the generation of a near-optimal sequence of motor selections. All

343   eleven endophenotypes show uniform distributions of action selections, complying with the

344   random distribution of the inputs configurations (**Fig. 3A**). This control condition allows the

345   simulated network to generate transient temporal responses that couple multiple initial states

346   with multiple stable states, in a transient winner-take-all or *winner-less* competition (Rabinovich

347   et al., 2006; Afraimovich et al., 2008).

348

349   Under the condition of simulated addiction, one of the actions is associated with drug

350   administration (**Fig. 3B**, values represented in blue). Substance use triggers phasic dopamine

351   bursts, leading to Hebbian learning in cortico-striatal connections of both dorsal and ventral

352   circuits (equation 3). In recurrent networks, circuit gain increases as a direct function of the

353   weights of reentrant synapses (Amit, 1989). A dopamine response triggered by healthy

354   unexpected rewards would create a bias towards the selection of the reinforced motor response to

355   a perceived cue (Cohen and Frank, 2009; Grahn et al., 2009). However, drug consumption

356   triggers extra-physiologic dopamine-dependent learning, resulting in aberrantly high circuit gain,

357   and compromising the ability of all affected circuits to discriminate among different inputs and

358   produce temporal transitions towards multiple stable states. The cortico-striatal circuits become

359   over-stable and resistant to perturbation caused by a change of input or by noise (Fiore et al.,

360   2014) as they are dominated by *parasitic attractors* (Hoffman and McGlashan, 2001) (**Fig 1C**).

361   In the ventral cortico-striatal circuit, a parasitic attractor sets and maintains the selection of drug-

362   related goals or outcomes, biasing the action-outcome assessments required for planning. In the

363   dorsal circuit, the same process determines over-stable selections of the reinforced motor

364   behavior, generating reactive responses and habits. Importantly, the learning process simulated in

365   our neural model led to the generation of parasitic attractors in both circuits across all

366   endophenotypes, as all agents eventually reached a fixed threshold in cortico-striatal neural

367   plasticity. Despite the generation of a form of compulsive drug seeking behavior across all

368   endophenotypes, we observed significant differences in motor response patterns as a function of

16

369   the balance between ventral and dorsal circuits. Specifically, the endophenotypes characterized

370   by unbalanced dorsal or ventral control (i.e. **Fig. 3B**, endophenotypes 1-3 and 9-11) expressed

371   distributions of motor selections that were significantly more compromised by drug-related

372   aberrant rewards, in comparison with balanced endophenotypes (i.e. **Fig. 3B**, endophenotypes 5-

373   7). The presence of identical learning processes, and the associated attractor formation in both

374   ventral and dorsal circuits, ascribes all phenotypic differences univocally to the only remaining

375   independent variable, which controls cortico-cortical connectivity and therefore the strength of

376   the biases between circuits. Unbalanced agents were characterized by more frequent drug-related

377   selections as actions leading to drug consumption were selected more frequently than in balanced

378   endophenotypes, in a range between +3% and +45%. This result would identify all phenotypes

379   within the limits of individual differentiation described in the study chosen for behavioral

380   comparison (Gannon et al., 2017).

381

382   Next, we investigated how the simulated endophenotypes behaved under the conditions of

383   treatment and relapse. We measured the frequency of drug-related action selections under the

384   conditions of addiction and treatment (**Fig. 4A-B**). Both ventral (goal-oriented) and dorsal

385   (habitual) treatments effectively reduced the number of actions associated with drug

386   consumption, in comparison with baseline addiction. However, the dorsal treatment was more

387   effective for dorsal-dominated endophenotypes and the ventral treatment was more effective for

388   ventral-dominated endophenotypes. These endophenotype-specific treatment effects were further

389   confirmed by our analysis of individual differences under the relapse condition (**Fig. 4C-D**):

390   dorsal treatments were more effective in elongating time to relapse for dorsal-dominated

391   endophenotypes, whereas ventral treatments were more successful in delaying relapse for

17

392   ventral-dominated endophenotypes. Together, both measures showed that our simulated

393   treatments have differential effects depending on the balance between dorsal and ventral circuits.

394   Importantly, these differences emerged only after the treatment was applied, where a pre-

395   treatment comparison between compulsive behaviors expressed by the opposite unbalanced

396   endophenotypes (i.e. ventral-dominant or dorsal-dominant) did not show any significant

397   difference in choice selections (cf. **Fig. 3B**, endophenotypes 1-3 and 9-11).

398

399   **Simulations from the RL model**.

400   By simulating explicit negative outcomes associated with drug consumption, the RL model

401   allowed to measure the likelihood each agent has to develop addiction, as a function of its

402   endophenotype. In our analysis, addiction is defined as a behavior leading to drug selections

403   more frequently than the healthy alternative reward, under the addiction phase. The mean

404   percentage of these *addicted* agents (over 300 runs) was 43.05%, across endophenotypes, which

405   is consistent with the percentage of rats developing compulsive self-administration of cocaine, as

406   reported in the reference study (~40% over a period of 5 days, cf. Gannon et al., 2017).

407   Importantly, when considering endophenotype differentiation, the percentage varies

408   significantly: 60.3% for $\beta = 0$, 40.3% for $\beta = 0.2$, 30.1% for $\beta = 0.4$, 36.7% for $\beta = 0.6$, 39.3% for

409    $\beta = 0.8$, and 51.6% for $\beta = 1$ (**Fig. 5A-B**). This phenotypic differentiation is consistent with well-

410   established data from animal models. For instance, rat strains selectively bred for either high or

411   low voluntary running differ in the likelihood to develop addiction when given free access to

412   cocaine (respectively ~35% and ~60% of each strain develop addiction over a period of 5 days,

413   cf. Smethells et al., 2016). Free access to substances of abuse does not necessarily lead to

414   compulsive behaviors (Piazza et al., 1989; Belin et al., 2011), as addiction varies as a function

18

415   factors such as exposure extent, amount of drug delivered, and associated negative effects

416   (Pelloux et al., 2007; Jonkman et al., 2012). Our simulations suggest that endophenotypes with

417   lower chances of addiction are characterized by balanced control modalities. Note that an

418   optimal agent, knowing the environment structure and being able to compute the long-term

419   effects of drug, will never select drug-states (**Table 3**).

420

421   Finally, the simulations suggest that the hypothetical treatment targeting model-free control is

422   the most effective, reducing the likelihood to pursue drug-related behaviors for all

423   endophenotypes (**Fig. 5A)**. In contrast, the model-based treatment appears to be less effective for

424   all endophenotypes, with the exception of the purely model-based one (β=1) (**Fig. 5B)**. Under

425   the condition of relapse, our data confirm that the simulated treatments significantly differ in

426   their effectiveness across the proposed endophenotypes, also suggesting the treatment targeting

427   model-free control is the most successful in prolonging relapse time (**Fig. 5C-D**). Relapse time

428   after model-free treatment is mostly similar to the time required to develop addiction behavior

429   before any treatment (**Fig. 5C**). At the opposite side of the control spectrum, the model-based

430   treatment shows a positive effect only for the purely model-based endophenotype. All remaining

431   endophenotypes show relapse times significantly shorter than those recorded the first time

432   addiction is developed (β=1; **Fig. 5D**).

433

434   **Discussion**

435   Individual differences in stress and anxiety responses (Dilleen et al., 2012; Jimenez and Grant,

436   2017), social dominance (Morgan et al., 2002; Covington and Miczek, 2005), aggressive

437   temperament (McClintick and Grant, 2016), preference for saccharine (Carroll et al., 2002),

438   sensation or novelty seeking (Suto et al., 2001; Nadal et al., 2002; Belin et al., 2011; Flagel et al.,

439   2014), impulsivity (Perry and Carroll, 2008; Verdejo-Garcia et al., 2008; Dalley et al., 2011),

440   and sensitivity to rewards (Belcher et al., 2014) have all been found in both animal models and

441   clinical studies in humans to be associated with addiction vulnerabilities, and in particular with

442   the likelihood to develop and maintain addiction, or to resist to treatment (Piazza et al., 1989;

443   Belin et al., 2016; Everitt and Robbins, 2016). However, investigations into the mechanisms

444   underlying this phenotypic differentiation in addiction has so far revealed few neural or

445   computational candidates, which are found associated with diverse and dissociable behavioral

446   traits. An important example is represented by the endophenotypic differentiation reported in the

447   expression and reactivity of striatal D2 dopaminergic receptors, which is found across the traits

448   of impulsivity (Dalley et al., 2007), social dominance (Morgan et al., 2002), novelty preference

449   (Flagel et al., 2014) and sensitivity to rewards (Belcher et al., 2014). The overlap of this

450   endophenotypic trait across multiple, non-coexisting, phenotypes associated with addiction

451   vulnerabilities suggests other neural or computational mechanisms have yet to be identified to

452   allow accounting for the reported variety in behavioural traits.

453

454   Here we have presented a neural field model, augmented by an RL model, to expand on existing

455   neuropsychological and computational accounts of addiction. Our models propose a theoretical

456   investigation into the interaction among cortico-striatal circuits or behavioral control modalities,

457   and the effects this interaction has on addiction development, and treatment response. As

458   classically described in previous models (Redish, 2004; Redish et al., 2008; Dayan, 2009), we

459   have assumed that over-evaluation of a drug leads to aberrant dopamine release and associated

460   over-learning in multiple DA targets (Bjorklund and Dunnett, 2007). In the neural field model,

461   this mechanism results in the dysregulation of the circuit gain and associated dynamics of both

462   ventral and dorsal cortico-striatal circuits. In the integrated model-based and model-free RL

463   model, sequential choice behavior is confounded by the presence of a high immediate reward

464   (drug state), which leads to misrepresent the negative outcomes following drug consumption, if

465   their distribution across states and time is sufficiently complex to escape the capabilities of the

466   agent to correctly represent the environment (Doll and Daw, 2016; Sadacca et al., 2016). We

467   found that both models jointly indicate that the balance between neural circuits or behavioral

468   control modalities is a candidate neurocomputational mechanism characterizing endophenotypes

469   in addiction. The neural and RL models converge in suggesting that individuals characterized by

470   balanced behavioral control between reward-seeking or planning (ventral circuit/model-based)

471   and reactive or habitual responses (dorsal circuit/model-free) would have a reduced chance to

472   develop addiction and decreased severity of symptoms if developing addiction. We propose this

473   neurocomputational mechanism may be interacting with other known endophenotypic

474   differentiations, such as alterations of D2 receptors in the striatum (Morgan et al., 2002; Nader

475   and Czoty, 2005; Dalley et al., 2007; Volkow et al., 2007; Belcher et al., 2014; Flagel et al.,

476   2014) or differences in learning rates (Gutkin et al., 2006; Piray et al., 2010), to generate the

477   multifaceted behavioral traits that have been reported in literature to be associated with addiction

478   vulnerabilities.

479

480   In our neural model, ventral and dorsal circuits are mostly in phase in their selections under the

481   pre-drug condition, exhibiting synchronous transient stability of neural activity and enhancing

482   the overall ability of the system to adapt to changing stimuli (i.e. the two circuits adapt to the

483   input changes with a similar pace and synchronize in their selection). Under the condition of

484   addition, the two circuits are mostly pulled towards the parasitic attractor state associated with

485   drug consumption, and they occasionally select the competing non-drug stimuli. If only one of

486   the two systems performs a selection outside of the attractor, the difference in selection generates

487   a dissonance or interference. In neural endophenotypes characterized by unbalanced control, this

488   dissonance is solved by one circuit taking the lead, so that both systems eventually converge on

489   the selection of the dominant circuit. These dynamics result in limited opportunities to generate

490   non-drug responses to the external stimuli, as they can only be generated by the dominant circuit.

491   Conversely, in balanced control endophenotypes, if any of the two circuits ignores the drug-

492   stimulus and selects a competing option, the resulting dissonance can trigger a state transition

493   pulling out the parasitic attractor states associated with substance use. The endophenotypes in

494   our simulations vary only in the parameters regulating the balance between circuits, as

495   dopamine-driven learning processes established between cortex and striatum (equation 3) do not

496   vary across endophenotypes, resulting in identical habit formation and drug-related biases in the

497   outcome representations. Thus, our proposed phenotypic differentiation does not interfere with

498   the usual role ascribed to the ventral and dorsal circuits as respectively implicated in the initial

499   reward-seeking phase in addiction (Belin and Everitt, 2008; Willuhn et al., 2012) and the

500   subsequent consolidation of stimulus-response, habitual, association (Everitt and Robbins, 2013,

501   2016). However, our simulated dynamics show that, after addiction is developed, systemic over-

502   stability can be reduced or further enhanced, depending on the cortico-cortical biases between

503   cortico-striatal circuits. In turn, this modulation of system stability can foster or further impair

504   input discrimination and motor response versatility, affecting addiction symptomatology. As a

505   result, our neural model shows phenotypic variability emerging after the presentation of the

506   reward simulating the drug and addiction is developed, in a gradient of over-selection of drug-

507   related actions.

508

509   In the RL model, we investigated whether the balance between model-based and model-free

510   modalities would also increase the robustness of the system against the selection of drug-states in

511   a more complex environment and in presence of explicit negative outcomes. Similar to the neural

512   model, a system with balanced control modalities introduces more diversity in action selection

513   during exploration, reducing (yet not cancelling) the chances of developing maladaptive reactive

514   responses. This increased diversity and overall reliability are likely to be induced by a higher

515   redundancy and diversification of the system. While both components may fail, the causes of

516   failures are not necessarily correlated. The model-based system can fail due to its sensitivity to

517   cognitive resources but it is more efficient in encoding previous experience of the agent. On the

518   other hand, the model-free component is more affected by limited exploration but it is more

519   reliable in its selections unaffected by the availability of cognitive resources. Consistent with the

520   neural model, differentiations in behaviors among endophenotypes emerge in an inverted-U

521   shape, where unbalanced control system are the most vulnerable to developing addiction.

522

523   The phenomenon of relapse is more elusive and the two models do not fully converge on this

524   aspect. In the neural model, balanced dorsal and ventral endophenotypes respond well to both

525   types of simulated treatments. For the unbalanced endophenotypes, however, only the

526   appropriate treatment, targeting the dominant neural circuit, is effective. The simulations in the

527   RL model do not show this same symmetric effects for the two treatments: the model-free

528   treatment is effective for most endophenotypes, whereas the model-based treatment is mostly

529   unsuccessful, with short relapse times across all endophenotypes, but the purely model-based

530   one. The latter result is possibly due to the model learning process characterizing the model-

531   based component, which is affected by conflicting information as drug use is associated with

532   both positive and negative outcomes, experienced by the agent when entering the drug state

533   under different conditions.

534

535   It is worth noting that habitual and goal-oriented behaviors have neural representations in the

536   dorsal and ventral cortico-striatal circuits respectively, but they do not fully overlap with model-

537   based and model-free control modalities in RL (Dolan and Dayan, 2013). Nonetheless, the neural

538   and RL models independently simulate choices among competing options in addiction. Thus, we

539   were able to test our hypothesis of endophenotypic differentiation under two complementary

540   levels in Marr's tri-level of analysis: the neural implementation and the algorithmic level (Marr

541   and Poggio, 1976). This multilevel modeling approach has been often used in computational

542   psychiatry (Maia and Frank, 2011; Montague et al., 2012; Adams et al., 2016; Hauser et al.,

543   2016; Huys et al., 2016) to highlight model convergence and associate specific neural structure

544   and dynamics with mathematical formalizations of optimal and suboptimal behavior in RL. The

545   convergence of neural and RL model also provides more confidence in the reliability of the

546   identified computational mechanisms underlying addiction and the associated characterization of

547   endophenotypes. Importantly, the simulated endophenotype-dependent treatment effects can

548   potentially inform the development of individualized treatments for addiction. For instance, both

549   models indicate individuals showing unbalanced cortico-striatal activity or control modality are

550   at higher risk of relapse and might need prolonged or ad hoc developed treatments, in

551   comparison with balanced endophenotypes. Further studies will be required to provide empirical

24

552   validation of our models. For example, connectivity analysis of fMRI imaging can be used to test

553   effective connectivity among cortico-striatal circuits, in conjunction with cognitive tasks

554   targeting the model-based and model-free control systems.

555

558

561

562

563

References

564

565   Adams RA, Huys QJ, Roiser JP (2016) Computational Psychiatry: towards a mathematically
566   informed understanding of mental illness. J Neurol Neurosurg Psychiatry 87:53-63.
567   Afraimovich V, Tristan I, Huerta R, Rabinovich MI (2008) Winnerless competition principle
568   and prediction of the transient dynamics in a Lotka-Volterra model. Chaos
569   18:043103.
570   Amit DJ (1989) Modeling brain function: the world of attractor neural networks. Cambridge
571   ; New York: Cambridge University Press.
572   Baldassarre G, Mannella F, Fiore VG, Redgrave P, Gurney K, Mirolli M (2013) Intrinsically
573   motivated action-outcome learning and goal-based action recall: a system-level bio-
574   constrained computational model. Neural Netw 41:168-187.
575   Balleine BW (2005) Neural bases of food-seeking: affect, arousal and reward in
576   corticostriatolimbic circuits. Physiol Behav 86:717-730.
577   Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control:
578   corticostriatal determinants of goal-directed and habitual action.
579   Neuropsychopharmacology 35:48-69.
580   Belcher AM, Volkow ND, Moeller FG, Ferre S (2014) Personality traits and vulnerability or
581   resilience to substance use disorders. Trends Cogn Sci 18:211-217.
582   Belin D, Everitt BJ (2008) Cocaine seeking habits depend upon dopamine-dependent serial
583   connectivity linking the ventral with the dorsal striatum. Neuron 57:432-441.
584   Belin D, Deroche-Gamonet V (2012) Responses to novelty and vulnerability to cocaine
585   addiction: contribution of a multi-symptomatic animal model. Cold Spring Harbor
586   perspectives in medicine 2.
587   Belin D, Belin-Rauscent A, Everitt BJ, Dalley JW (2016) In search of predictive
588   endophenotypes in addiction: insights from preclinical research. Genes, brain, and
589   behavior 15:74-88.
590   Belin D, Mar AC, Dalley JW, Robbins TW, Everitt BJ (2008) High impulsivity predicts the
591   switch to compulsive cocaine-taking. Science 320:1352-1355.
592   Belin D, Berson N, Balado E, Piazza PV, Deroche-Gamonet V (2011) High-novelty-
593   preference rats are predisposed to compulsive cocaine self-administration.
594   Neuropsychopharmacology 36:569-579.
595   Bjorklund A, Dunnett SB (2007) Dopamine neuron systems in the brain: an update. Trends
596   Neurosci 30:194-202.
597   Carroll ME, Morgan AD, Lynch WJ, Campbell UC, Dess NK (2002) Intravenous cocaine and
598   heroin self-administration in rats selectively bred for differential saccharin intake:
599   phenotype and sex differences. Psychopharmacology (Berl) 161:304-313.
600   Cohen MX, Frank MJ (2009) Neurocomputational models of basal ganglia function in
601   learning, memory and choice. Behavioural brain research 199:141-156.
602   Covington HE, 3rd, Miczek KA (2005) Intense cocaine self-administration after episodic
603   social defeat stress, but not after aggressive behavior: dissociation from
604   corticosterone activation. Psychopharmacology (Berl) 183:331-340.
605   Dalley JW, Everitt BJ, Robbins TW (2011) Impulsivity, compulsivity, and top-down
606   cognitive control. Neuron 69:680-694.
607   Dalley JW, Fryer TD, Brichard L, Robinson ES, Theobald DE, Laane K, Pena Y, Murphy ER,
608   Shah Y, Probst K, Abakumova I, Aigbirhio FI, Richards HK, Hong Y, Baron JC, Everitt

609   BJ, Robbins TW (2007) Nucleus accumbens D2/3 receptors predict trait impulsivity
610   and cocaine reinforcement. Science 315:1267-1270.
611 Daw ND, Dayan P (2014) The algorithmic anatomy of model-based evaluation. Philos Trans
612   R Soc Lond B Biol Sci 369.
613 Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and
614   dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704-1711.
615 Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on
616   humans' choices and striatal prediction errors. Neuron 69:1204-1215.
617 Dayan P (2009) Dopamine, reinforcement learning, and addiction. Pharmacopsychiatry 42
618   Suppl 1:S56-65.
619 Deco G, Jirsa VK, Robinson PA, Breakspear M, Friston K (2008) The dynamic brain: from
620   spiking neurons to neural masses and cortical fields. PLoS Comput Biol 4:e1000092.
621 desJardins ME, Durfee EH, Ortiz J, Charles L., Wolverton MJ (1999) A survey of research in
622   distributed, continual planning. AI Magazine 20:13-22.
623 Dilleen R, Pelloux Y, Mar AC, Molander A, Robbins TW, Everitt BJ, Dalley JW, Belin D (2012)
624   High anxiety is a predisposing endophenotype for loss of control over cocaine, but
625   not heroin, self-administration in rats. Psychopharmacology (Berl) 222:89-97.
626 Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80:312-325.
627 Doll BB, Daw ND (2016) The expanding role of dopamine. Elife 5.
628 Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement
629   learning: a behavioral and neurocomputational investigation. Brain Res 1299:74-94.
630 Draganski B, Kherif F, Kloppel S, Cook PA, Alexander DC, Parker GJ, Deichmann R,
631   Ashburner J, Frackowiak RS (2008) Evidence for segregated and integrative
632   connectivity patterns in the human Basal Ganglia. J Neurosci 28:7143-7152.
633 Economidou D, Pelloux Y, Robbins TW, Dalley JW, Everitt BJ (2009) High impulsivity
634   predicts relapse to cocaine-seeking after punishment-induced abstinence. Biol
635   Psychiatry 65:851-856.
636 Ersche KD, Turton AJ, Pradhan S, Bullmore ET, Robbins TW (2010) Drug addiction
637   endophenotypes: impulsive versus sensation-seeking personality traits. Biol
638   Psychiatry 68:770-773.
639 Everitt BJ, Robbins TW (2013) From the ventral to the dorsal striatum: devolving views of
640   their roles in drug addiction. Neurosci Biobehav Rev 37:1946-1954.
641 Everitt BJ, Robbins TW (2016) Drug Addiction: Updating Actions to Habits to Compulsions
642   Ten Years On. Annu Rev Psychol 67:23-50.
643 Fiore VG, Dolan RJ, Strausfeld NJ, Hirth F (2015) Evolutionarily conserved mechanisms for
644   the selection and maintenance of behavioural activity. Philos Trans R Soc Lond B
645   Biol Sci 370.
646 Fiore VG, Rigoli F, Stenner MP, Zaehle T, Hirth F, Heinze HJ, Dolan RJ (2016) Changing
647   pattern in the basal ganglia: motor switching under reduced dopaminergic drive. Sci
648   Rep 6:23327.
649 Fiore VG, Sperati V, Mannella F, Mirolli M, Gurney K, Friston K, Dolan RJ, Baldassarre G
650   (2014) Keep focussing: striatal dopamine multiple functions resolved in a single
651   mechanism tested in a simulated humanoid robot. Front Psychol 5:124.
652 Flagel SB, Waselus M, Clinton SM, Watson SJ, Akil H (2014) Antecedents and consequences
653   of drug abuse in rats selectively bred for high and low response to novelty.
654   Neuropharmacology 76 Pt B:425-436.

655 Flagel SB, Robinson TE, Clark JJ, Clinton SM, Watson SJ, Seeman P, Phillips PE, Akil H (2010)
656   An animal model of genetic vulnerability to behavioral disinhibition and
657   responsiveness to reward-related cues: implications for addiction.
658   Neuropsychopharmacology 35:388-400.
659 Gannon BM, Galindo KI, Rice KC, Collins GT (2017) Individual Differences in the Relative
660   Reinforcing Effects of 3,4-Methylenedioxypyrovalerone under Fixed and
661   Progressive Ratio Schedules of Reinforcement in Rats. The Journal of pharmacology
662   and experimental therapeutics 361:181-189.
663 Garrison KA, Potenza MN (2014) Neuroimaging and biomarkers in addiction treatment.
664   Current psychiatry reports 16:513.
665 Gerfen CR, Surmeier DJ (2011) Modulation of striatal projection systems by dopamine.
666   Annu Rev Neurosci 34:441-466.
667 Gershman SJ, Horvitz EJ, Tenenbaum JB (2015) Computational rationality: A converging
668   paradigm for intelligence in brains, minds, and machines. Science 349:273-278.
669 Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND (2016) Characterizing a psychiatric
670   symptom dimension related to deficits in goal-directed control. Elife 5.
671 Gould RW, Duke AN, Nader MA (2014) PET studies in nonhuman primate models of cocaine
672   abuse: translational research related to vulnerability and neuroadaptations.
673   Neuropharmacology 84:138-151.
674 Grahn JA, Parkinson JA, Owen AM (2009) The role of the basal ganglia in learning and
675   memory: neuropsychological studies. Behavioural brain research 199:53-60.
676 Gruber AJ, McDonald RJ (2012) Context, emotion, and the strategic pursuit of goals:
677   interactions among multiple brain systems controlling motivated behavior. Front
678   Behav Neurosci 6:50.
679 Gutkin BS, Dehaene S, Changeux JP (2006) A neurocomputational hypothesis for nicotine
680   addiction. Proc Natl Acad Sci U S A 103:1106-1111.
681 Haber S (2008) Parallel and integrative processing through the Basal Ganglia reward
682   circuit: lessons from addiction. Biol Psychiatry 64:173-174.
683 Haber SN (2003) The primate basal ganglia: parallel and integrative networks. J Chem
684   Neuroanat 26:317-330.
685 Hauser TU, Fiore VG, Moutoussis M, Dolan RJ (2016) Computational Psychiatry of ADHD:
686   Neural Gain Impairments across Marrian Levels of Analysis. Trends Neurosci 39:63-
687   73.
688 Hoffman RE, McGlashan TH (2001) Neural network models of schizophrenia.
689   Neuroscientist 7:441-454.
690 Huys QJ, Maia TV, Frank MJ (2016) Computational psychiatry as a bridge from
691   neuroscience to clinical applications. Nat Neurosci 19:404-413.
692 Hyman SE, Malenka RC, Nestler EJ (2006) Neural mechanisms of addiction: the role of
693   reward-related learning and memory. Annu Rev Neurosci 29:565-598.
694 Jahanshahi M, Obeso I, Rothwell JC, Obeso JA (2015) A fronto-striato-subthalamic-pallidal
695   network for goal-directed and habitual inhibition. Nat Rev Neurosci 16:719-732.
696 Jimenez VA, Grant KA (2017) Studies using macaque monkeys to address excessive alcohol
697   drinking and stress interactions. Neuropharmacology 122:127-135.
698 Jonkman S, Pelloux Y, Everitt BJ (2012) Drug intake is sufficient, but conditioning is not
699   necessary for the emergence of compulsive cocaine seeking after extended self-
700   administration. Neuropsychopharmacology 37:1612-1619.

701 Jupp B, Dalley JW (2014) Behavioral endophenotypes of drug addiction: Etiological insights
702   from neuroimaging studies. Neuropharmacology 76 Pt B:487-497.
703 Keramati M, Dezfouli A, Piray P (2011) Speed/accuracy trade-off between the habitual and
704   the goal-directed processes. PLoS Comput Biol 7:e1002055.
705 Koob GF, Volkow ND (2016) Neurobiology of addiction: a neurocircuitry analysis. The
706   lancet Psychiatry 3:760-773.
707 Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and
708   neurological disorders. Nat Neurosci 14:154-162.
709 Marr D, Poggio T (1976) From understanding computation to understanding neural
710   circuitry. Cambridge: Massachusetts Institute of Technology, Artificial Intelligence
711   Laboratory.
712 McClintick MN, Grant KA (2016) Aggressive temperament predicts ethanol self-
713   administration in late adolescent male and female rhesus macaques.
714   Psychopharmacology (Berl) 233:3965-3976.
715 Molander AC, Mar A, Norbury A, Steventon S, Moreno M, Caprioli D, Theobald DE, Belin D,
716   Everitt BJ, Robbins TW, Dalley JW (2011) High impulsivity predicting vulnerability
717   to cocaine addiction in rats: some relationship with novelty preference but not
718   novelty reactivity, anxiety or stress. Psychopharmacology (Berl) 215:721-731.
719 Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn
720   Sci 16:72-80.
721 Moore A, Atkeson CG (1993) Prioritized sweeping: Reinforcement learning with less data
722   and less time. Machine Learning 13:103-130.
723 Morgan D, Grant KA, Gage HD, Mach RH, Kaplan JR, Prioleau O, Nader SH, Buchheimer N,
724   Ehrenkaufer RL, Nader MA (2002) Social dominance in monkeys: dopamine D2
725   receptors and cocaine self-administration. Nat Neurosci 5:169-174.
726 Nadal R, Armario A, Janak PH (2002) Positive relationship between activity in a novel
727   environment and operant ethanol self-administration in rats. Psychopharmacology
728   (Berl) 162:333-338.
729 Nader MA, Czoty PW (2005) PET imaging of dopamine D2 receptors in monkey models of
730   cocaine abuse: genetic predisposition versus environmental modulation. Am J
731   Psychiatry 162:1473-1482.
732 Nestler EJ, Aghajanian GK (1997) Molecular and cellular basis of addiction. Science 278:58-
733   63.
734 Obeso JA, Rodriguez-Oroz MC, Stamelou M, Bhatia KP, Burn DJ (2014) The expanding
735   universe of disorders of the basal ganglia. Lancet 384:523-531.
736 Pelloux Y, Everitt BJ, Dickinson A (2007) Compulsive drug seeking by rats under
737   punishment: effects of drug taking history. Psychopharmacology (Berl) 194:127-
738   137.
739 Pelloux Y, Murray JE, Everitt BJ (2015) Differential vulnerability to the punishment of
740   cocaine related behaviours: effects of locus of punishment, cocaine taking history
741   and alternative reinforcer availability. Psychopharmacology (Berl) 232:125-134.
742 Perry JL, Carroll ME (2008) The role of impulsive behavior in drug abuse.
743   Psychopharmacology (Berl) 200:1-26.
744 Pezzulo G, Rigoli F, Chersi F (2013) The mixed instrumental controller: using value of
745   information to combine habitual choice and mental simulation. Front Psychol 4:92.

746   Piazza PV, Deminiere JM, Le Moal M, Simon H (1989) Factors that predict individual
747   vulnerability to amphetamine self-administration. Science 245:1511-1513.
748   Piray P, Keramati MM, Dezfouli A, Lucas C, Mokri A (2010) Individual differences in nucleus
749   accumbens dopamine receptors predict development of addiction-like behavior: a
750   computational approach. Neural Comput 22:2334-2368.
751   Rabinovich MI, Varona P, Selverston AI, Abarbanel HDI (2006) Dynamical principles in
752   neuroscience. Reviews of Modern Physics 78.
753   Redish AD (2004) Addiction as a computational process gone awry. Science 306:1944-
754   1947.
755   Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in
756   the decision process. Behav Brain Sci 31:415-437; discussion 437-487.
757   Sadacca BF, Jones JL, Schoenbaum G (2016) Midbrain dopamine neurons compute inferred
758   and cached value prediction errors in a common framework. Elife 5.
759   Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward.
760   Science 275:1593-1599.
761   Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J,
762   Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J,
763   Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T,
764   Hassabis D (2016) Mastering the game of Go with deep neural networks and tree
765   search. Nature 529:484-489.
766   Simon DA, Daw ND (2012) Dual-system learning models and drugs of abuse. In:
767   Computational Neuroscience of Drug Addiction, 1 Edition (Gutkin B, Ahmed SH,
768   eds), pp 145–161. New York: Springer-Verlag.
769   Singh S, Jaakkola T, Littman ML, Szepesvári C (2000) Convergence Results for Single-Step
770   On-Policy Reinforcement-Learning Algorithms. Machine Learning 38:287-308.
771   Smethells JR, Zlebnik NE, Miller DK, Will MJ, Booth F, Carroll ME (2016) Cocaine self-
772   administration and reinstatement in female rats selectively bred for high and low
773   voluntary running. Drug and alcohol dependence 167:163-168.
774   Suto N, Austin JD, Vezina P (2001) Locomotor response to novelty predicts a rat's
775   propensity to self-administer nicotine. Psychopharmacology (Berl) 158:175-180.
776   Sutton RS (1990) Integrated architecture for learning, planning, and reacting based on
777   approximating dynamic programming. In: Proceedings of the seventh international
778   conference (1990) on Machine learning, pp 216-224. Austin, Texas, USA: Morgan
779   Kaufmann Publishers Inc.
780   Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT
781   Press.
782   Verdejo-Garcia A, Lawrence AJ, Clark L (2008) Impulsivity as a vulnerability marker for
783   substance-use disorders: review of findings from high-risk research, problem
784   gamblers and genetic association studies. Neurosci Biobehav Rev 32:777-810.
785   Volkow ND, Morales M (2015) The Brain on Drugs: From Reward to Addiction. Cell
786   162:712-725.
787   Volkow ND, Fowler JS, Wang GJ, Swanson JM, Telang F (2007) Dopamine in drug abuse and
788   addiction: results of imaging studies and treatment implications. Archives of
789   neurology 64:1575-1579.
790   Voon V, Reiter A, Sebold M, Groman S (2017) Model-Based Control in Dimensional
791   Psychiatry. Biol Psychiatry 82:391-400.

792   Watkins CJ, Dayan P (1992) Q-Learning. Machine Learning 8:279-292.
793   Willuhn I, Burgeno LM, Everitt BJ, Phillips PE (2012) Hierarchical recruitment of phasic
794   dopamine signaling in the striatum during the progression of cocaine use. Proc Natl
795   Acad Sci U S A 109:20703-20708.
796   Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve
797   outcome expectancy but disrupt habit formation in instrumental learning. Eur J
798   Neurosci 19:181-189.
799

800

801

802   **Figure 1. Illustrative representation of energy landscapes and neural architecture of the**

803   **model. A-C.** These representations of energy landscapes are meant to illustrate differences in the

804   temporal responses provided by neural systems. Depending on the energy landscape, three

805   arbitrary inputs (magenta circles) are transformed into different stable states (grey circles).

806   Learning processes increase or decrease the strength of the connections among nodes in a

807   network, thereby altering its energy landscape and reshaping temporal responses towards

808   existing attractors. Attractors are defined as low energy states (bottom of the basins) at the end

809   point of the temporal responses to multiple starting inputs. **A.** The landscape is characterized by

810   multiple shallow attractors: these allow slow temporal responses, transforming multiple inputs

811   into multiple weakly stable states. Noise and changes in the incoming input easily determine new

812   responses towards different attractors. **B.** In this second illustrative configuration, steep and vast

813   attractors characterize the energy landscape, allowing quick state transitions towards two

814   equilibrium points. This new configuration is able to resist noise and minor changes in the

815   incoming input, and, at the same time, allows a differentiation of inputs in two broad categories.

816   **C.** Finally, the third energy landscape illustrates the presence of a parasitic attractor,

817   exemplifying the condition of addiction: all inputs fall now at the bottom of a single steep basin.

818   Under this condition, noise and changes in the incoming input determine temporal responses that

819   keep falling in the same attractor, therefore preventing the system from executing different

820   behaviors. **D.** Neural architecture used to simulate neural dynamics and behavior for the mean

821   field neural model. The activity in the dorsal cortico-striatal circuit is responsible for the motor

822   output of the system (left circuit), whilst activity in the ventral cortico-striatal circuit is

823   responsible for goal selections (right circuit). The two systems bias each other via cortico-

824   cortical connectivity and learning processes affect the weights of the connections between the

825   two cortical outputs and the striatum in their corresponding circuits. The components in the

826   architecture are labeled as follows: cortex (Cx), thalamus (Th), globus pallidus pars externa and

827   interna (GPe and GPi), substantia nigra pars reticulata (SNr), sub-thalamic nucleus (STN) and

828   striatum (Str), divided into two areas enriched by either D1 or D2 receptors.

829

830   **Figure 2**. Illustrative representation of the environment used for the RL model of addiction. The

831   states are disposed in a linear arrangement: on one extreme is a healthy reward state (1), on the

832   opposite side a drug state (8) followed by fourteen aftereffects states (9-22). Healthy reward and

833   drug states are separated by 6 neutral states (2-7). The agent can traverse between nearby neutral

834   states. From the two borders of the central segment of neutral states, an agent can enter the

835   healthy reward state (from state 2), securing a moderate reward (Rg=1), or the drug state (from

836   state 7), receiving an initial high reward (Rd=10, under the condition of addiction) and a series of

837   sparse but temporally extended negative outcomes, characterizing the aftereffects states. The

838   presence of negative outcomes makes entering the drug and aftereffects area suboptimal under

839   all experimental conditions (see optimal policy in **table 3**). From both the goal state and the

840   drug/aftereffects segment the agent is then returned to the middle of the neutral segment. In this

841   representation, we explicitly portray the transitions related to states 1 (healthy reward), 4

842   (neutral), and 15 and 20 (drug aftereffects) for illustrative purposes. Line width represents related

843   transition probability value. Line and text color represent the action class (a_s, a_g, a_w, a_d). Neutral

844   states are navigable with actions a_{s2-7} which are deterministic for adjacent state while have high

845   chance of failing for distant states. From the neutral states the agent can reach: (i) the healthy

846   reward, if executing action a_g when in state 2; and (ii) the drug state (8) and aftereffects area

847   (state 9 to 22), if executing action a_d, when in state 7. From the healthy reward area the agent can
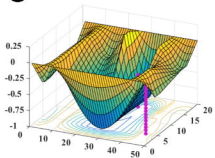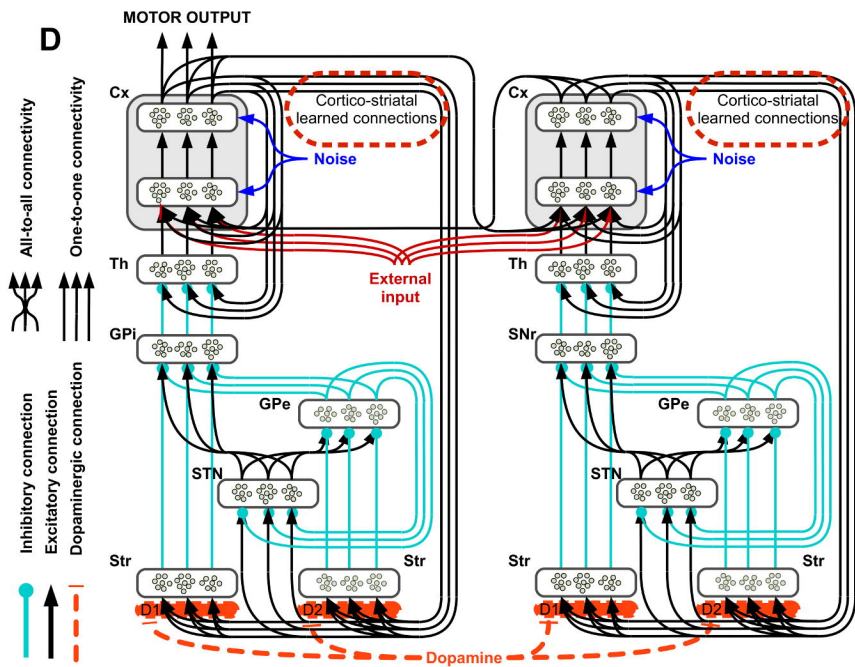
848   issue again a_g, receiving a reward of 1 and going back to the center of the neutral area, state 4.

849   By entering the drug area, the agent receives a reward of 10. Action results in the drug/aftereffect

850   area are probabilistic: the agent can reach a nearby state in the area or leave the area and reach

851   the center of the neutral state. Leaving the drug/aftereffects area has a cost of -4, whereas every

852   other transition inside the area costs -1.2. For a full description of transitions and their

853   probability distribution in the environment (see **Tables 1-2,4-5**).

854

855

856   **Figure 3.** Histograms show how the distribution of simulated action selections changes

857   depending on the endophenotype (11 structural variations in cortico-cortical connectivity). 30

858   random seeds/inputs are used per phenotype, tested under two conditions: pre-drug (A) and

859   addiction (B). The three colors represent the occurrence of selections of three arbitrary actions.

860   Under the pre-drug condition, no reward is provided and action selections are triggered by

861   random fluctuation in values of competing sensory inputs. The simulations show the agents adapt

862   to the changes in sensory stimuli and therefore exhibit a near-uniform distribution of action

863   selections. Conversely, under the condition of addiction, the action represented in blue is

864   associated with administration of the simulated drug, triggering DA-dependent Hebbian learning

865   and consequently over-selection. Under addiction, the differences among endophenotypes clearly

866   emerge in the selection frequency of the action leading to drug consumption. Asymmetric control

867   (endophenotypes 1-3 and 9-11) leads to a stronger over-selection in comparison with balanced

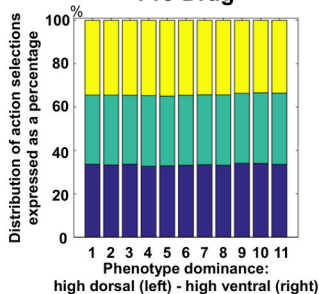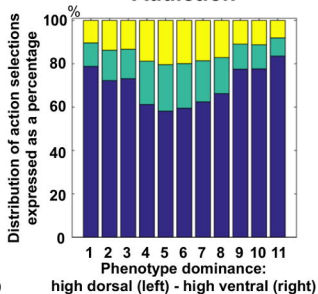868   control (endophenotypes 4-7), despite identical learning processes and reward encoding.

869

870

34

871 **Figure 4.** Shaded error bars report mean and standard error for 30 simulated agents across 11

872 endophenotypes (structural variations in cortico-cortical connectivity). Panels A and B (top

873 panels) show the selections of actions leading to substance consumption, as a percentage of the

874 overall number of action selections. In the first case (A) we compare the values recorded under

875 the condition of addiction with those recorded under the condition of dorsal treatment jointly

876 with abstinence (i.e. drug-related actions do not trigger self-administration of a drug and the

877 treatment targets the dorsal circuit). In the second case (B) the comparison involves addiction

878 and ventral treatment (treatment targeting the ventral circuit, during abstinence). Panels C and D

879 (bottom panels) compare the simulated time required by the 11 endophenotypes to reach an

880 arbitrary threshold of cortico-striatal connectivity under the condition of addiction and under the

881 condition of relapse after either dorsal (C) or ventral (D) treatment. Within the time of a

882 simulation run, all simulated agents reached the addiction threshold under all conditions here

883 reported. The two treatments are simulated by restoring either the dorsal/motor (A-C) or the

884 ventral/outcome circuit (B-D) to the configuration characterizing the pre-drug condition. The

885 percentage of the action selections shows the dorsal treatment is more effective in

886 endophenotypes characterized by high dorsal dominance (A), whereas the ventral treatment only

887 has an effect in endophenotypes characterized by high ventral dominance (B). Similarly, dorsal

888 and ventral treatments result in long relapse times in endophenotypes characterized by high

889 dorsal and high ventral dominance, respectively. (*) indicates significant difference: $p < 0.05$.

890

891 **Figure 5.** Shaded error bars report mean and standard error for □100 simulated agents across 6

892 endophenotypes (differential balance between model-based and model-free control modalities,

893 β=[0, 0.2, 0.4, 0.8, 1]). Panels A and B show the percentage of agents developing addiction (i.e.

894   drug-related choices are more frequent than healthy reward-related choices), per endophenotype,

895   under the condition of addiction and treatment. In the first case (A) the comparison involves data

896   recorded under the condition of addiction and those recorded under the condition of model-free

897   treatment. In the second case (B) the comparison involves the condition of addiction and model-

898   based treatment. Panels C and D illustrate the simulated time required by the 6 endophenotypes

899   to reach 95% of action preference towards the drug state, in comparison with action preference

900   recorded under the condition of addiction (f2). In the first case (C) the comparison involves the

901   conditions of addiction and relapse after model-free treatment, whereas in the second case (D)

902   the comparison involves the conditions of addiction and relapse after model-based treatment. In

903   terms of action selection ratio, the simulated results show both treatments have a significant

904   effect only on those phenotypes characterized by strong unbalance of control (A-B). In terms of

905   relapse, the results show the model-free treatment is on average more successful than the model-

906   based one, as 5 endophenotypes show no significant difference between the condition of

907   addiction and post-treatment addiction (i.e. the time required to relapse is not significantly

908   different than the time required to develop addiction the first time). Each endophenotype, or

909   parameter selection, was simulated 100 times across the four phases (3050 steps per simulation).

910   Results depend on the statics of the environment, but over similar environments the results were

911   qualitatively similar. (*) indicates significant difference: $p<0.05$.

A Pre-Drug
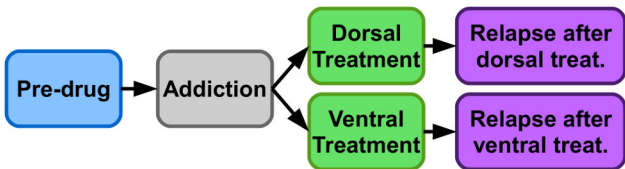
B Addiction

Phenotype dominance:
high dorsal (left) - high ventral (right)

Distribution of action selections expressed as a percentage
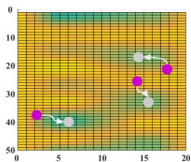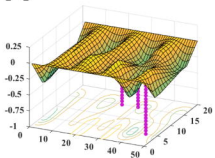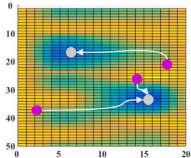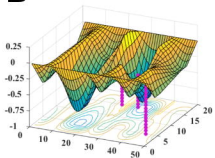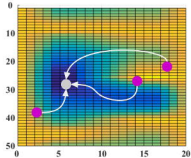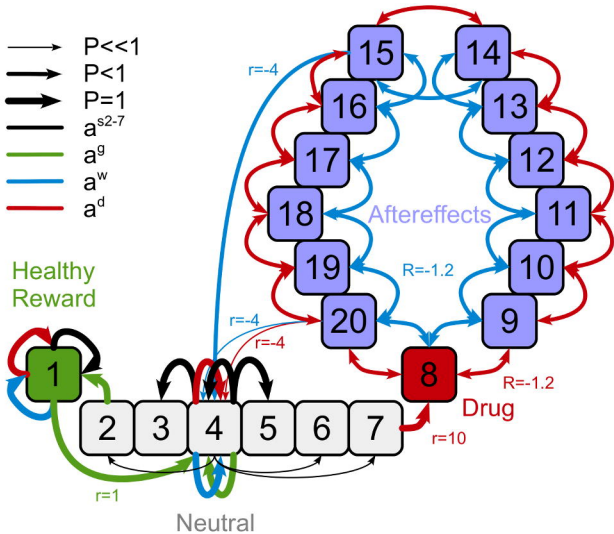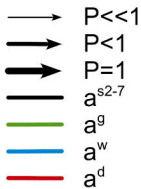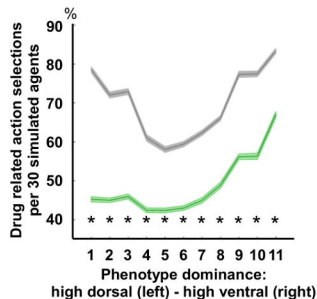
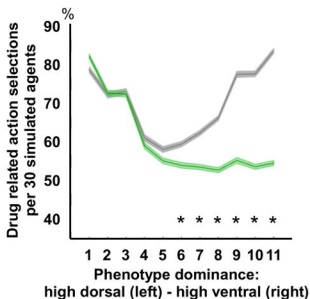Action 1, never reinforced
Action 2, never reinforced
Action 3, reinforced with addictive substance, under addiction condition

A
Dorsal Treatment
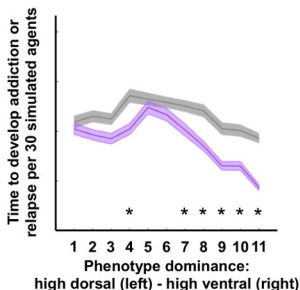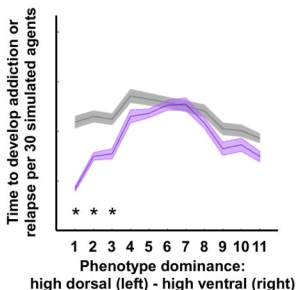
B
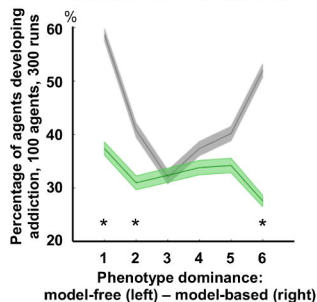**Ventral Treatment**
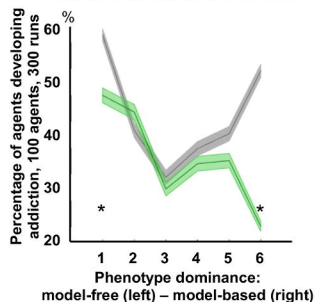
C
Dorsal Treatment

D
Ventral Treatment