Dopamine, Reinforcement Learning, and Addiction

Section 2 very briefly sketches a modern theory of neural RL [23], pointing out the diverse influences it accords to dopamine. Section 3 considers dopamine's role in the early stages of addiction. Section 4 considers possible extensions to computations, including the one due to [81] involving saturating prediction errors, one concerned with satiating action propensities (8) or boosted advantages [5], and two sorts of Pavlovian responses [109, 27]. Although we attempt to build an integrative account, it is important to remember that there are many very important differences between different drugs of addiction; further, it is not presently generally possible to capture all the complex effects at multiple temporal scales over the release and reception of dopamine itself and other neuromodulators, and also over other aspects of systems involved in control. Further, in keeping with the special issue, we focus particularly on dopamine, leaving many other issues associated with RL models of addiction to [23] and the extensive critical commentary associated with that paper.

Reinforcement learning and dopamine

At a computational level [73, 11], RL offers theories of learning to predict and act appropriately in affectively charged, partially unknown, environments [118]. In most interesting cases, the environment has multiple states (like locations in a maze), with actions causing stochastic but well-sampled transitions between states and, perhaps occasionally, giving rise to desirable or undesirable reinforcement outcomes such as foods, drugs or electrical shocks. RL is often considered in instrumental or operant terms with the subject having at least partial agency. However, in the case that the subject never has a choice (which one might think of as there is only one option), exactly the same computations methods allow the learning of action-effect predictions, and future outcomes, which is normally the preserve of Pavlovian or classical conditioning. We mainly discuss RL in the richer, operant case, but refer to Pavlovian issues as they arise. In environments such as a maze, an action cannot only be judged by its immediate consequences; rather, it is necessary to consider the cumulative utility of all the outcomes arising in the future that depend on the action. This makes for a computationally challenging problem. RL includes different algorithmic approaches to this challenge. Notably, one recent model-based and model-free methods [20]. In turn, these have rather different neural implementations [6], some, but not others, of which critically involve dopamine.

In model-based RL, subjects are assumed to build so-called forward models of their environments. These specify the probabilities that particular outcomes or state transitions arise from particular actions, and also report the utilities of those actions. Optimal choices in model-based RL is conceptually simple, involving forward or backwards search in the tree of all the accessible states to find the actions leading to the largest cumulative reward. It is also straightforward to handle uncertainty correctly, training off exploration for exploitation [46, 21]. However, this conceptually simple approach is not what is usually bought at what is typically a hugely computationally intensive price for searching the tree, or alternatively a proportionally large computational advantage induced by searching the tree. Indeed, this is the price of doing the computation. Since model-based decisions are made on the basis of predictions of actual outcomes, they can automatically be sensitive to the utilities of those outcomes that apply to the subject's current motivational state. In psychological terms, model-based RL is goal-directed (i.e., animals choose actions because they expect particular, desired, outcomes to result [31]). There is evidence in rats that some aspects of goal-directed control, notably subliminal and, not dependent on the assumption that voluntary movement is a whole is diminished by dopamineergic deficits [74, 83], perhaps by its effects on the nucleus accumbens, where it is thought to promote action. In model-free RL, subjects acquire ways of evaluating or predicting the long-term sum of utilities associated with executing actions, without building up any formal form of forward model. Versions of these utilities include what are known as Q values [123] and advantages [5]. They can be learned in the absence of a model on the basis of the fact that predictions of long run utility should be consistent across alternative action trajectories. For instance, an action at one stage has a higher value if it leads directly to a high utility outcome, or leads to a transition to a second stage with a high value that itself is valued.

Any inconsistency gives rise to a prediction error that can be used to correct the value of the initial state. Of course, early in learning, the value of the second state will not be accurate, and so this form of "bootstrapping" is statistically inefficient. Nevertheless, Q and advantage values are simple to use, since they automatically obviate the need for search, with actions associated with larger predicted utilities being selected more frequently. Since the predictions are the sum of the utilities of the ultimate outcomes, and outcomes themselves, model-free control is insensitive to the current motivational state of the subject. In psychological terms, model-free RL is habitual [31]. In the end, the relative utilities of different outcomes are important; it is only necessary to have a model of the best at each stage. Thus, there is a spectrum of model-free RL methods, with the two poles of the model-free spectrum defined by the utility of the action at the stage of the choice process and the prediction error that leads to the reward or utility.

Dopamine plays a substantial role in model-free RL, in both Pavlovian [107] and instrumental [79, 102] settings, with evidence that its phasic activity [108] and release [22] represents aspects of the prediction error, and, in instrumental settings, a more general role as an "error signal." This error signal is used to update the reinforcement learning weights. Dopamine activity is thus used to update the weights of the reinforcement learning model, which in turn is used to predict the reward associated with a given action. This process is repeated over time, allowing the animal to learn the optimal action for each situation. Over time, the animal learns to associate dopamine activity with the likelihood of reward, and uses this information to guide its behavior. Dopamine activity is thus an essential component of learning in reward-based environments.
conceived slightly differently there [51], for prediction errors at a more ventral part of the spiral (acting as the critic) to teach action values encoded at a more dorsal part (the actor). Dopamine plays at least two further roles in modern theories of RL. First, there is an association between tonic levels of the neuromodulator (which may be partially independent of phasic release [47]) and the value of energy at responding [84]. This has been interpreted in RL terms as arising from the additional degree of freedom of choosing the latency of executing an action in order to balance the excess energetic cost of acting very quickly against the opportunity cost of missing out on potentially available rewards by acting very slowly. [84] suggested that tonic dopamine reports the average rate of (consolable) reward. This acts as an acquisition level — states or actions associated with rates of reward lower than the current average will be relatively more reinforced. In temporal terms, this average is exactly the opportunity cost of time, and via the trace described above, is positively correlated with vigor. [84] discussed the account this provides of the data implicating an involvement of dopamine in effort costs [103, 106]. It could also relate to the psychomotor activating properties of dopamine-boosting stimulants (which is itself a venerable idea in addiction: [110]), and increased impulsivity, since the higher the opportunity cost, the greater the price of a delay, and the less willing subjects will be to wait for rewards.

The second role concerns a phenomenon called Pavlovian to instrumental transfer [71]. In this, subjects are separately trained on two contingencies, an instrumental one such as lever pressing for food, and a Pavlovian one, such as the association between a conditioned stimulus such as a tone and another reward (B). If the subject is then allowed to press the lever expecting reward A being provided, then it will press the lever expecting both food and the tone also being provided (also without the actual delivery of reward B). The greatest win regards comes if the Pavlovian and instrumental outcomes are linearly identical (a circumstantial case for (8)) but lever pressing is enhanced even if the Pavlovian outcome is different (general [97]), providing that its current motivational value is positive (so that the stimulus is experienced as an effect and not reward unless the subject is thirsty). PT appears to depend on Pavlovian values, positively represented in the amygdala [15], affecting the nucleus accumbens, and it is mapped by drugs that boost dopamine in the accumbens [133]. One idea (general [97]) consistent with this dopaminergic influence is that precessing the presentation of the expected reward will result in higher levels of dopamine at its targets [103, 114, 120]. The simplest model in which dopamine is a pure appetitive influence, in which there would be just the attraction rate that is associated with the predictions rather than acting to increase the value or propensity of states or actions (since there is no prediction error to start with, there would be no dopamine to be subject to offset hiccough), however, there is also a baseline or tonic release of dopamine [47], and if opioids boost this, they would create an imaginary reward that would have the same effects as above [111]. There is also evidence that novel stimuli and states lead to phasic dopamine activity [88]; an effect modelled in RL prediction error terms as being a spur to exploration [66]. This could provide an initial phasic dopamine response on which the opioids would subsequently have their effects.

Finally, there is ample evidence that opioids enhance the specific motivational functions of certain brain regions such as the SNC [98, 88], and may have various aversive outcomes less unpleasantly [38]. Assuming something like a baseline level of activation of the systems involved in evolved environments (perhaps as part of an opponent dopamine responses, it is to be presumed that these responses are adaptive in environmentally relevant environments; and indeed experiments have come to use them in working out what instrumental actions to make on the basis of learning in experiments. Nevertheless, instrumentally motivated actions such as approach or withdrawal of the face of a reinforcer or their coming into the picture of the predictors, could, for instance, lead to maladaptive outcomes such as impulsivity or learning effects in choice

**Initiation**

We first consider the dopaminergic processes that are initially engaged by drugs of addiction and that lead them, if sampled, to be likely to be repeated. One obvious route, lies within model-free RL. We argued that the phasic activity of dopaminergic cells acts as an error associated with predictions of future reward. When this is positive, this implies that more reward than expected has been provided. In turn, dopaminergically-controlled plasticity should increase the original prediction that then appears erroneously pessimistic. If this prediction is associated with a state, then this would make that state more attractive, potentially leading to Pavlovian approach and other effects such as conditioned place preference [110]. If the prediction is the Q value or advantage of a particular action at a state, then increasing this will increase the frequency with which that action will be chosen at the state [51]. Thus, drugs that cause dopamine concentrations to be higher at key pyramidal target, by blocking reserpine, fructose, releasing it from transporter, or by voltaged-dependent feedback inhibition, or directly increasing the phase activity of dopaminergic neurons, should lead to some of the first signs that drugs can act as (positive) reinforcers: increased an interest in the model-free actor–critic instrumental conditioning architecture, the phasic release of dopamine criticizes the choice of action. In this case, drug-induced increases will inflate the propensity to perform the associated action. The apparently subtle difference between this case and the case of value and advantage is discussed below in terms of the evolution of compulsions.

Along with these direct routes, there are also some possibilities for indirect influence. For instance, opioids can act to magnify the effect of dopamine at its targets [103, 114, 120]. In the simplest model in which dopamine is a pure appetitive influence, this would actually just change the learning rate associated with the predictions rather than acting to increase the value or propensity of states or actions (since there is no prediction error to start with, there would be no dopamine to be subject to offset hiccough), however, there is also a baseline or tonic release of dopamine [47], and if opioids boost this, they would create an imaginary reward that would have the same effects as above [111]. There is also evidence that novel stimuli and states lead to phasic dopamine activity [88]; an effect modelled in RL prediction error terms as being a spur to exploration [66]. This could provide an initial phasic dopamine response on which the opioids would subsequently have their effects.

**Structure:** [100] provides an additional route by which model-free RL could be affected. Importantly, by this means, opioids, and other drugs working on specific evaluation systems, could also influence model-based RL [100]. Since the involvement of dopamine in goal-directed control is believed to be relatively constrained, this offers one route towards non-dopaminergic aspects of addiction [66, 67].

**Compulsion**

Compared with this relatively restricted set of ideas about the indirect influences of drug-taking, the long-run behavioral and neural effects of addictive drugs appear to be more heterogenous than between different substances, and exhibit substantial variation between different individuals. They have also led to a number of other different theoretical ideas (well aired in the literature and associated commentary). The direct relevance of RL concepts (let alone the influence of dopamine), is somewhat questionable, that is to underlie many of the criticisms expressed to that paper. However, one of the main aspects of maladaptive decision making in addiction is the evolution of the compulsive consummatory use of drugs, i.e., that they are sought and consumed despite evident knowledge of their negative consequences (e.g., [33, 71]). In this section, we discuss some of the key candidate RL-based routes by which dopamine’s influence on at least the first stages of compulsive behavior might be explained. Before focusing on the role that dopamine might play, we need first to understand what the structure of a compulsive behavior might be in the context of RL. Normally, the long-run costs of actions are assumed to be weighed together with their benefits to give rise to Q-values advantage action or action evinements. Given values for two such actions (QQ, QD), the second action (B) has a smaller cost. However, not all drugs have a similar effect, and a more likely the larger the value of (QQ, QD), Fig. 1A shows this signal. The general idea is to consider how being an unusually strongly held impetus towards an action associated with the delivery of a drug. The first two routes to compulsion come from problems with incorrectly specified action values or expected rewards. The difference is that these basic action-value queries are correct, but that they fail to determine decision-making completely, for instance because of the long-term influence of Pavlovian responses over instrumental behavior. We discuss two routes to compulsion from this.

It is first important to see the likely insufficiency of the simple mechanism we have discussed above by which drug-induced dopamine signals can interact to initiate responses. It is easiest to do this by writing down the basic RL equations (which, in this simpli-
The equation representing the relationship between the Q value and the action value is:

\[ Q = \frac{1}{1 + e^{-a(x-b)}} \]

where \( a \) and \( b \) are parameters, and \( x \) is the input value.

The learning rate \( \alpha \) is used to update the Q value after each action is taken.

\[ \Delta Q = \alpha (R + \gamma \max Q' - Q) \]

where \( R \) is the reward, \( \gamma \) is the discount factor, and \( Q' \) is the next state Q value.

The Q-learning algorithm updates the Q value based on the following equation:

\[ Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max Q'(s', a') - Q(s, a)] \]

where \( Q(s, a) \) is the Q value for state \( s \) and action \( a \), \( r \) is the immediate reward, \( \gamma \) is the discount factor, and \( Q'(s', a') \) is the Q value for the next state \( s' \) and action \( a' \).

This algorithm allows the agent to learn the optimal policy by iteratively updating the Q values based on the observed rewards and the expected future rewards.
over goal-directed control (which is not), then they can have the effect of weakening the will, and so enhancing the compulsive effects. Indeed, these drug-induced deficits in regions of prefrontal cortex (which is associated with goal-directed control) both in animal models of addiction and human addicts themselves (see [34,123]). Theories of the balance between the systems that regulate reward and those that control it have been developed in a heuristic manner by the relative certainties of the systems rather than their relative predictions [20]. The means by which the certainties are calculated (although not necessarily regulated in a heuristic manner) are dependent on the neurochemical systems that are involved in the dopamine system, which are involved in reward prediction, the ventral tegmental area, and the nucleus accumbens. The hallmark of addiction is a failure to suppress the dopamine system when it is activated, leading to an increase in craving and the pursuit of drug use. This is because the reward system is hypersensitive to the reinforcing effects of drugs, leading to a decrease in the threshold for reward and an increase in the response to rewarding stimuli.

Addiction is typically characterized by repeated cycles of abuse, abstinence, and relapse [86,113]. It is hard to provide a definitive RL account of relapse, because of the complexities of the extinction processes that are presumably happening during abstinence. Relapse occurring as a result of reinforcement processes to drugs themselves or not directly associated with drugs is relatively straightforward to accommodate; however, additional factors over relapse such as stress, and the effect of stress on the response of dopamine systems to drugs or cues [85,113], is more complex in being dependent on the interaction of drug craving and environmental cues. It is likely that the relationship between stress and drug craving is complex, and that stress can have both positive and negative effects on drug craving. The stress response is complex and involves a network of stress-related systems and neurotransmitters, including the hypothalamic-pituitary-adrenal axis, the sympathetic-nervous system, and the immune system. The stress response can be facilitated by drugs, leading to an increase in drug craving and relapse. Therefore, it is important to understand the mechanisms underlying stress and drug craving, and the effects of stress on drug craving and relapse.

References
