Word Count: 4235

Appendix: 1347

Chapter 12:
Toward Precision Cognitive Behavioral Therapy via Reinforcement Learning Theory

Yael Niv, PhD
Professor; Co-Director, Rutgers-Princeton Center for Computational Neuropsychiatry (CCNP), Princeton Neuroscience Institute & Psychology Department, Princeton University, Princeton, NJ, USA
yael@princeton.edu

Peter Hitchcock, PhD
Postdoctoral Research Associate, Laboratory of Neural Computation and Cognition, Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI, USA
peter_hitchcock@brown.edu

Isabel M. Berwian, [PhD – almost!]
Postdoctoral Research Associate, Translational Neuromodeling Unit, University of Zurich and ETH Zurich, Zurich, Switzerland
berwian@biomed.ee.ethz.ch

Gila Schoen, MD
Psychiatrist, Child and Adolescent Division, Geha Mental Health Center, Petah Tikva, Israel
schoen-g@zahav.net.il

Corresponding author: Yael Niv, PhD, Princeton Neuroscience Institute Room 143, Washington Rd, Princeton, NJ 08540

Chapter Headings

Highlights

- The goal of cognitive behavioral therapy (CBT), a widely used method for many psychiatric conditions with a strong empirical evidence base, is to effect change in cognition, emotion, and behavior.
- Concomitant with the development of CBT, cognitive neuroscientists and psychologists converged on a theoretical framework called reinforcement learning (RL) to explain trial-and-error learning, behavioral decision-making, and their neural implementation. RL suggests that learning occurs and expectations (values) are updated when we experience a "prediction error" – a discrepancy between what we expected and what actually happened.
- Humans employ both model-free and model-based decision-making mechanisms. The former depends on dopaminergic prediction-error-based learning, the ventral and dorsolateral striatum and portions of the amygdala, and reflects ingrained, less flexible, habitual behavior. The latter seems dopamine-independent, relies on the hippocampus, frontal cortex, and the dorsomedial striatum, and reflects flexible, prospective goal-directed planning.
- The CBT framework can be mapped directly to core concepts from RL. The cognitive aspects of CBT may be seen as targeting the model-based system, while the behavioral component of CBT is strongly related to the model-free learning system.
- RL may aid in improving CBT through the development of more effective treatment protocols, quantification of individual parameters of the learning process to help better tailor CBT, and measurement of the effects of psychotherapy.

Introduction: Learning as a Basis for Mental Health Treatment

Cognitive behavioral therapy (CBT; Beck AT 1976; Beck AT 2005), a leading method of psychotherapy, is a widely used treatment for many psychiatric conditions including anxiety disorders, mood disorders, post-traumatic stress disorder, obsessive-compulsive disorder, substance abuse and more. The goal of CBT, in particular the second wave of CBT on which we will focus (Beck JS 2011), is to effect change in cognition, emotion, and behavior, understanding that changing any one of these affects change in the other two due to their interdependence. This is done by teaching patients new thought ("cognitive") and response ("behavioral") patterns with the aim of eventually, through repeated practice and generalization, displacing maladaptive automatic tendencies. A key CBT principle is that emotional responses (e.g., feeling scared, anxious or sad) are not a direct consequence of external events, but rather are mediated by thoughts and internal schemas that filter external information and determine emotional and behavioral responses. By learning and practicing alternative interpretations and responses, we can therefore gain control over our emotional wellbeing. CBT has been widely tested in randomized control trials and has a strong empirical evidence base (Butler et al 2006; Hofmann et al 2012).

Over the same time frame, research in cognitive neuroscience and psychology has converged on a theoretical framework called reinforcement learning (RL; Sutton and Barto 2018) to explain trial-and-error learning, behavioral decision making, and their neural implementation. The key idea is that choices are based on learned *values* of available options. These values reflect the subjective reward value of expected outcomes (which could be negative for aversive outcomes) and are learned through trial and error, by comparing actual outcomes to expectations. The principle is that when we experience a "prediction error" – a discrepancy between what we expected and what actually happened – learning occurs and expectations (values) are updated.

RL theory has been widely tested in humans and animals, accumulating much support. Recent advances have extended the theory to explain multiple learning algorithms, and their respective decision processes, which occur in parallel in the brain (in particular, habitual versus goal-directed deliberative behavior; Daw, Niv, Dayan 2005), and to thinking about how learning is generalized or specialized to different scenarios (Gershman, Norman, Niv 2015). This enhanced RL framework can be mapped relatively directly onto key concepts from CBT.

The goal of this chapter is to flesh out the links between second-wave CBT and RL[1]. Because RL theory is defined in computational terms (see below), viewing CBT through the lens of RL can suggest ways to quantify the changes that a patient undergoes throughout treatment, and to potentially help determine which CBT intervention to employ at each stage of treatment. This highlights how developments in the growing field of "computational psychiatry" (Huys, Maia, Frank 2016; Montague et al 2012) could impact clinical practice.

Reinforcement Learning in a Nutshell

RL arose as a theory of animal learning, specifically Pavlovian and instrumental conditioning (Sutton and Barto 1990). In *Pavlovian ("classical") conditioning*, a contingency between two events, one motivationally neutral and one motivationally relevant (e.g., the sound of a bell—a "conditional stimulus," or CS—followed by receipt of food, an "unconditional stimulus" or US) is experienced repeatedly. As a result, through learning, the CS comes to *predict* the occurrence of the US. This prediction is evidenced (and can be measured) by behavioral responses (e.g., salivation, quickening heart rate – a host of "conditioned responses") that automatically accompany said prediction. This type of learning is ubiquitous and can occur when a CS predicts the occurrence of a US (excitatory conditioning) or the absence of an otherwise available US (inhibitory conditioning). Importantly, learning may not enter awareness, and the automatic conditioned responses (which include emotions, increased heart rate, sweating) are very hard to override.

This type of prediction learning is at the heart of RL theory (for a more detailed overview of RL in the brain, see Niv 2009). In RL, each stimulus *(S)* acquires a value *V(S)* that reflects the subjective scalar value of the USs it predicts. For example, the first time the bell is heard, its value may be zero. If the tone is followed by a US, the US's reward value *R* (positive values for appetitive USs like food, and negative values for aversive USs like pain) is compared to the prior expectation $V_{old}(S)$ to compute a prediction error, $PE = R - V_{old}(S)$. The new value of the stimulus is then updated based on the prediction error: $V_{new}(S) = V_{old}(S) + \alpha \cdot PE$, where α is a "learning rate" parameter between 0 and 1. This learning rule will update the value every time a prediction error is experienced, until the prediction error is zero (the prediction is correct). This

[1] Due to space limitations, we forgo discussion of the links between CBT and Bayesian inference processes (which are strongly linked to RL), and refer the reader instead to Moutoussis et al. (2017).

model of trial-and-error prediction learning has one parameter, α, that can differ between individual learners or for different situations (see appendix A1.1).

When the environment changes considerably (e.g., a tone CS that once reliably predicted a shock US is no longer followed by shock), instead of updating $V(S)$, the learner may infer that the current tone is not the same as the old one, and thus initialize a completely new value $V(S_2)$. The idea is that the learner makes inferences about the hidden (latent) causes of observed events, and learns different values for different inferred latent causes. Rather than thinking of S as a stimulus, it denotes an inferred *state* of the world, corresponding to a latent cause (Gershman, Blei, Niv 2010).

In *instrumental conditioning*, actions that the learner chooses can affect whether they will or will not receive reinforcement. For example, pressing a lever may be reinforced with food, or with removal of an aversive sound. The learner must experiment with different actions and learn, through trial and error, which actions are effective in any given scenario. In RL terminology, the model for learning is similar to the prediction learning model; however, values are learned for each possible action a in each state (these values are called Q-values and denoted $Q(a|S)$, in contrast to the V-values for states, above). Again, learning proceeds based on prediction errors: after performing a chosen action in the current state, the outcome is used to compute a prediction error and update $Q(a|S)$.

This learning mechanism is only one way to compute values. Rather than storing and updating values after each experience, one can instead learn a *model of the environment*: how states follow one another given each action (the state "transition structure"), and what states are rewarding or punishing. Armed with this model, the learner can mentally simulate the consequences of different actions (or, in Pavlovian scenarios where actions are irrelevant, the unfolding of events over time) to calculate their value. This alternative algorithm has been termed "model-based learning," in contrast to the "model-free" trial-and-error learning algorithm described above (Daw, Niv, Dayan 2005).

It appears that the brain uses both model-free and model-based methods for computing values and making decisions (Daw, Niv, Dayan 2005). The model-free system depends on dopaminergic prediction errors and a cortico-basal-ganglia-thalamocortical loop involving the ventral and dorsolateral striatum and portions of the amygdala (see appendix A1.2). It has been associated with *habitual* behavior – well learned response patterns that require considerable experience with conflicting outcomes to learn new values and new actions. In contrast, the model-based system seems dopamine-independent, and relies on the hippocampus, frontal cortex, and a cortico-basal-ganglia-thalamocortical loop involving the dorsomedial striatum. This system has been associated with deliberative, so-called *goal-directed* behavior, and is more flexible in its action selection as it can incorporate new information into the task model without extensive experience. However, simulating future outcomes requires mental effort, and there may be limits to how deeply one can search a tree of future options. Neural and behavioral evidence show that both systems operate in the brain in parallel; however, one or the other may be controlling behavior at any point in time. For example, you may use model-based RL to explicitly plan to go shopping on the way home from work. However, listening to the radio as you drive, your model-free system may take over and lead you to turn towards home habitually despite your plan otherwise.

In sum, RL theory suggests that to change behavioral responses, one can change the internal model of the task, which will affect deliberative planning (see also Moutoussis et al. 2017). However, since much of our behavior is habitual and relies on a lifetime of learning from trial and error, more permanent change may require experiencing prediction errors that will slowly change the values that our model-free system assigns to states and actions. Pharmacological treatments that affect dopamine (commonly used for a variety of mental health conditions) can affect this latter process as it depends on dopaminergic prediction errors. Moreover, if a situation changes too much, the learner may infer a new state rather than update an old state value. In psychotherapy, such new inferences may be helpful as long as future experience is ascribed to the newly inferred latent cause. Yet, because the mental representation of the old latent cause remains, quiescent but unchanged, an associated maladaptive judgment or behavior can reappear if the individual infers that this cause has returned. In the examples below, we will flesh out the implications of these ideas for psychotherapy.

Cognitive Behavioral Therapy in a Nutshell

CBT is a problem-focused collaborative form of psychotherapy that aims to change maladaptive behavior and thought processes, and improve emotional regulation (Beck JS 2011). A core premise of CBT is that external events do not cause us to feel and do things, but rather our cognitions offer a *subjective interpretation* of events which, in turn, causes feelings and actions. This interpretation is often automatic and implicit, building on a lifetime of previous experiences, and not a voluntary or conscious process. The profound implication is that we have some control over our emotional, cognitive and behavioral responses. By changing our interpretations, we can avoid responding maladaptively (for an example, see Appendix 2.1).

How can interpretations be challenged and changed? Since emotions, thoughts and actions are inextricably linked, the therapist can choose which avenue may be most amenable to change for each individual. For example, actions and emotional responses can result from thoughts. In the method of *cognitive restructuring* these thoughts are challenged, their exaggerated or distorted nature is exposed, and alternatives are listed and practiced. This can help reduce the emotional response and enable alternative behavioral responses. Alternatively, *exposures* to seemingly dangerous (but actually safe) situations and experiencing their neutral outcomes can help reduce maladaptive automatic emotional responses (such as the autonomic fear response) that arise as “false alarms”. Once the emotional and bodily response is identified as a false alarm, one can additionally learn to turn off the alarm through relaxation and mindfulness techniques, or to wait it out knowing that there is no actual danger involved (for more detail, see Appendix A2.2). Moreover, reduction of the physiological stress itself enables increased flexibility of thought and action.

Cognitive Behavioral Theory in Reinforcement Learning Terminology

The CBT framework for understanding and treating dysfunction can be mapped quite directly to core concepts from RL (for mapping to the related framework of Bayesian inference, see Moutoussis et al. 2017). In RL terms, idiosyncratic response patterns emanate from state and action values learned from direct experience (or from modeling by others, e.g., our parents) through model-free trial and error, or computed on the fly from a learned world model.

At a first pass, the *cognitive* aspects of CBT (e.g., cognitive restructuring) can be seen as targeting the *model-based* system. Cognitive distortions may manifest in or result from distortions in the learned (or assumed) model of the environment. For example, if one’s estimated probability of transitioning from any state to a state accompanied by punishment is exaggerated, then every plan of action in this model may seem dangerous, leading to avoidance behavior. This mapping of model-based RL to the cognitive aspects of CBT suggests several avenues for change, each exploited in CBT: one can attempt to change the model of the environment, targeting the distorted transition estimates by challenging the validity of current assumptions, or by forcing oneself to execute a response plan that is estimated to produce negative outcomes and experiencing that these do not occur. This latter planned exposure will lead to changes both in the model of the environment and in learned action values used by the model-free system. In this way, both systems will promote healthier response patterns in the future.

Indeed, the *behavioral* component of CBT is strongly related to the *model-free* learning system: by orchestrating experiences and exposures that will lead to prediction errors, predictive values of states and actions can be re-trained. Importantly, the model-free system may be (at least partly) inaccessible to cognitive methods – we cannot talk our amygdala and striatum into changing stored values without having experienced prediction errors. Indeed, a hallmark of model-free learning is that avoidance of relevant training experiences serves to maintain old values. Thus, after experiencing a traumatic event (e.g., a car accident), the more one avoids similar scenarios (e.g., by not driving) the longer the negative expected values will be maintained. They may even become engrained over time, and potentially generalized to other (no longer experienced) scenarios (although RL theory does not currently model this phenomenon). This may explain why exposure-based methods are critical for CBT, and cognitive restructuring cannot usually stand alone. For an RL account of why exposure methods work, see Appendix A2.3.

Of note, RL theory and experimental findings suggest that model-based and model-free RL can occur even without direct experience, for example, by watching others behave and through mental simulation of internal models or replay of memories of previous experiences (Burke et al. 2010; Shohamy and Daw 2015). This may be a mechanism for runaway exaggerated values. Counterfactual learning from imaginary "what if" scenarios based on extreme and erroneous beliefs could lead to vastly distorted values, whereas real-world experience would better ground values in reality.

In this sense, the model at the heart of model-based learning (and related training of model-free values using internal simulations) may be the main source of value distortions – this system relies more directly on sampling from episodic memory, where outlier (e.g., traumatic) events are strongly encoded and preferentially retrieved (Brown and Kulik 1977; Madan et al. 2014; Rouhani and Niv 2019). This could lead to a distorted model that must be corrected through direct experience tied to the actual statistics of events in the environment, rather than their distorted representation in memory.

From Theory to Practice

CBT practice provides different protocols for different diagnoses and underlying pathologies. Here we briefly discuss two examples to illustrate the link to RL theory, and point out how theoretical work in RL can further our understanding of the mechanisms of CBT treatment and development of better protocols (see Craske et al. 2014, for a similar approach using animal learning theory). We will then discuss the implications of these links to precision psychiatry, and how quantification of an individual's learning and decision-making parameters can help tailor the therapeutic approach to each specific patient.

Example: Prolonged Exposure for Treatment of Post-Traumatic Stress Disorder

One of the best tested and most effective CBT methods for treating post-traumatic stress disorder (PTSD) is *prolonged exposure* (PE; Foa 2011). In PE, during *imaginal exposures,* the patient retells the story of the traumatic event in the first-person and present tense, and in as much detail as possible (mentioning all senses – vision, smell, etc.) in the safety of the clinic. The story is recorded, and the patient listens to it daily. In each of 6-8 therapy sessions, the patient retells the story (sometimes uncovering previously forgotten details) and receives the recording to listen to at home. Over time, putatively through a combination of desensitization/ habituation and re-organization of the memory through retelling, the traumatic memory becomes less potent and the patient recovers function: trauma-related thoughts become less intrusive, the patient no longer feels under constant threat, and aided by *in vivo* exposures, behaviors that had been avoided after the trauma are resumed.

The theory behind PE suggests that for the protocol to be successful, two important conditions must be met: the retelling of the trauma has to gain access to the original fear construct (i.e., the original fear memory) and disconfirming evidence then has to be introduced (Foa and Kozak 1986). Although developed separately (with PE far predating the relevant RL theory; Gershman et al. 2017), this method is extremely well aligned with the RL playbook: disconfirming experiences of safety generate prediction errors, and access to the original fear construct ensures that new learning is applied to the original state and not to a new state. Indeed, the protocol can cause much distress in the beginning – the patient is requested to revisit in detail memories that they have (unsuccessfully) tried to suppress for months and years, and to do so in an immersive way. If they are willing to do this, however, the method is very effective (Powers et al. 2010).

The precise and quantitative nature of formal RL theory may help us improve exposure therapy, especially where current recommendations are in conflict with experimental and theoretical work in RL. For example, building on inhibitory learning theory, Craske et al. (2014) recommend maximizing "expectancy violation" (the discrepancy between a patient's predicted and actual experience, i.e., the prediction error) during exposure to maximize learning of new safety associations that will compete with old trauma-related associations. However, the existence of multiple learning systems (model-based and model-free) suggests a more nuanced interpretation of some of their empirical results, and alternative methods for treatments. For example, one can modify the old association, that is, update previously learned values. Here, recent research suggests that large prediction errors can cause a learner to impute a new experience to a wholly different state, instead of revaluing an old state, thus leaving the original state value intact (Gershman et al. 2014; Gershman, Norman, Niv 2015). Indeed, experiments in rodents have shown that in the long-term, "gradual extinction" of a fear memory is more effective in modifying conditioned fear responding than abrupt extinction (Gershman et al. 2013), which suggests that moderate prediction errors are more effective than large ones. The use of RL theory to understand the principles that underlie when prediction errors maximally overwrite existing learning vs. are relegated to new states, and measuring individual differences in such thresholds using behavioral tasks (see Summary), may eventually help predict what type of exposure will be most effective for each patient.

An important caveat is that Pavlovian fear conditioning – the dominant animal model for PTSD and the experimental paradigm discussed above – assumes that learning is of a simple association between stimuli and an unavoidable aversive outcome. A traumatic memory is clearly more complex than a simple CS-US association, and more research is needed regarding how emotional events influence the organization of episodic memories (Cohen & Kahana, 2019; Talmi et al. 2019). Similarly, our understanding of PTSD could be increased through measuring individual differences in learning from positive versus negative prediction errors (Arkadir et al. 2016) and work on generalization, and specifically why negative memories tend to be stronger and generalize more widely than positive memories. More generally, understanding how boundaries between different situations (states) are drawn by the brain, and how these change over time (with and without experience of similar states), would be especially informative for treating PTSD.

Example: Exposure and Response Prevention in Obsessive-Compulsive Disorder

Obsessive-compulsive disorder (OCD) is characterized by obsessions (thoughts, e.g., “there are deadly germs on my hands that may make me very ill”) that increase subjective distress and anxiety, and compulsive actions that are performed to reduce this distress (e.g., washing hands, sometimes repeatedly). A prominent CBT treatment of OCD is *exposure and response prevention* (EXRP; Abramowitz 1996; Meyer 1966), wherein the patient is guided through a series of exposures to sources of distress while avoiding the action that would reduce the distress. The goal is for the patient to learn that 1) the distress subsides over time even without performing the compulsion (so they can forgo the compulsion), and 2) the terrible outcome that they thought would happen if the compulsion is not performed does not occur (they don’t die of a deadly disease).

From an RL perspective, one can conceptualize OCD in terms of Q-values – the values of different actions in different states. In OCD, the values of many states become negative *unless* a specific action is executed. So while the Q-value a=“washing hands” is 0 for many states, the Q-value of doing nothing in these states is presumably very negative. Performance of the obsession confirms the zero Q-value of the obsessive action (as nothing bad happens), yet prevents experiencing the outcome of not performing the obsession, so the value of other alternatives can remain erroneously negative. EXRP provides these learning experiences, which, through prediction errors, can retrain the negative values to 0. Moreover, because the value of a=“doing nothing” may generalize more widely to the value of the state in general, this training may generalize and prevent other obsessive actions (although, as mentioned, research on generalization in RL is in its infancy).

The EXRP protocol also includes imaginal exposures, in which the patient is asked to imagine and write down the worst-case scenario outcome of not performing the obsession. This challenges traditional extinction-based theories of OCD treatment, because there is no expectancy violation in this method. Rather, the patient is asked to imagine exactly what they fear will happen. RL theory may help explain why this is helpful: being forced to explicitly state the contingencies leading to the worst-case outcome may elaborate an otherwise sparsely-represented world model such that the exact series of events will be clearly represented, together with their low transition probabilities. This explicitly-thought-through representation may thereby decrease the estimated probability of feared outcomes.

Summary: Precision Psychotherapy – How Can Reinforcement Learning Theory Help?

CBT has its roots in behaviorism and decades of experiments on animal learning. Harnessing ideas from the contemporary version of this rich body of knowledge – reinforcement learning theory – can help develop even more effective treatment protocols. For example, RL experiments have shown that people have a higher learning rate for actions they choose freely compared to those they are forced to execute (Cockburn, Collins, Frank 2014). This suggests that offering several options for actions in each situation may speed skill learning in CBT, and may explain why it is beneficial to involve patients in the planning of exposures.

The quantitative nature of RL may also benefit precision psychiatry. The theory defines the dynamics of learning as a set of equations, and therefore allows quantification of individual parameters of the learning process using simple computerized decision-making tasks (Daw 2011). Indeed, fitting RL models to trial-by-trial choices in simple laboratory tasks is a means of measuring parameters such as an individual's learning rate (Niv et al. 2012), how this learning rate adapts to change in the environment (Behrens et al. 2007), differences in the rate of learning from positive versus negative prediction errors (Arkadir et al. 2016), the initial value ascribed to new choice options (Wittmann et al. 2008), and the tendency to use model-free versus model-based values in decision making (Gillan et al. 2016). Much current research in the field of *computational psychiatry* attempts to relate these parameters to psychopathologies, and to discover individual differences that may be transdiagnostically linked to mental illness (Bennett, Silverstein, Niv 2019; Huys, Maia, Frank 2016; Montague et al. 2012). The hope is that measuring quantities that relate to the mechanisms underlying psychiatric illnesses in an objective way that does not rely on self-report will allow more precise treatment predictions.

Applied to CBT, this approach can also assist in measuring individual parameters of learning that can help a therapist better target the treatment methods, and track progress over time. For example, knowing that a patient's model-free learning is slower than their model-based learning can help set the pace for exposures relative to cognitive restructuring methods, or assist in determining which method would be more effective. Moreover, overall impaired trial-and-error learning can suggest that CBT will not be an effective method for this patient, and perhaps pharmacological treatment should be the first line of action. In this way, we may make progress on the long-sought-after goal of assigning patients to the most effective treatment (Cohen 2018).

Finally, parameters from RL models may be useful for measuring the effects of psychotherapy. As examples, a decrease in the tendency to assign negative values to novel stimuli (Wittmann et al. 2008) may reflect reduction in over-generalizing negative prior knowledge, and increased flexibility in making choices (i.e., how willing a person is to explore options that do not have the highest value; Wilson et al. 2014) may track improvement due to treatment. Arguably, the goal of CBT is to provide patients with more flexible response options. By tracking flexibility using a simple computerized task administered repeatedly throughout the course of treatment, one can potentially determine when improvements have reached asymptote. In this way, RL theory – together with the set of tasks used to measure parameters of the learning process – can help develop *precision psychotherapy,* a therapeutic approach that uses objectively measurable indices of learning to most effectively help an individual.

Appendix 1: Reinforcement Learning

A1.1 Understanding Learning Rates in Reinforcement Learning

Another way to write the RL update equation $V_{new}(S) = V_{old}(S) + \alpha \cdot (R - V_{old}(S))$ is $V_{new}(S) = V_{old}(S) \cdot (1 - \alpha) + \alpha \cdot R$. This form highlights that learning is a weighted average between old knowledge $V_{old}(S)$ and new experience R, with the learning rate α determining the weighting of old and new information: high learning rates prioritize new experiences and cause the effects of old events to be "forgotten," whereas low learning rates allow values to reflect the effect of more past events. We emphasize that there is no single "correct" learning rate – in a stable but noisy environment it is advantageous to average over many events, whereas following an abrupt change it makes sense to learn quickly from new events (Yu and Dayan 2005). Indeed, experiments show that humans adjust their learning rate across tasks, and even within a task, in response to volatility as well as other factors (McGuire et al. 2014; Nassar et al. 2012). Therefore, even this extremely simplified one-parameter model can be used to describe interesting behavior that (normatively) adapts to task demands, and may potentially be disrupted due to mental illness.

A1.2 A Circuit for Model-Free Reinforcement Learning

Neurally, dopamine signaling is widely believed to correspond to RL prediction errors (Barto 1995; Montague, Dayan, Sejnowski 1996; Schultz, Dayan, Montague 1997). Numerous studies have shown that phasic dopamine bursts or pauses appear at times in a task where, in theory, the animal should be experiencing a prediction error. In humans, functional neuroimaging studies have identified a similar signal in the blood-oxygenation activity recorded in the ventral striatum (Hare et al. 2008) – an area that receives dense dopaminergic projections. Indeed, the striatum, which receives widespread projections from sensory, motor and associative cortical areas, is thought to be the area representing the values of states and actions. Learning in cortico-striatal synapses is modulated by dopamine, with dopamine concentration determining whether long-term changes in synapses will strengthen the synapse (long-term potentiation; when there is a surge of dopamine above baseline) or weaken it (long-term depression; when there is a dip in dopamine concentration below baseline) (Reynolds et al. 2001). This mechanism can easily implement the above-described trial-and-error learning algorithm, as positive prediction errors signaled by phasic increases in dopamine concentration would lead to more firing of striatal neurons in the presence of the state or the state and action in the future (signaling the now-higher expected value) and vice-versa for negative prediction errors (dips in dopamine firing).

Appendix 2: Cognitive Behavioral Therapy

A2.1 Example application of CBT principles to an event

Imagine someone interrupts you in a discussion. This event can raise different thoughts and interpretations, ranging from "how annoying, X is always so inconsiderate" to "what I was saying was probably not interesting...I bet everyone was relieved when he changed the topic". Each will lead to a different emotional response (e.g., feeling anger, insult, self-doubt) and different behavioral responses (e.g., try to speak over the interrupter; fall silent and speak less in this group in the future). According to CBT theory, individuals have internal *schemas* – ingrained beliefs that "filter" incoming information, biasing its subjective interpretation. For example, the schema "I am not good at anything" will favor the interpretation that one's contribution was boring or incorrect, whereas the schema "the world is against me" may favor the interpretation that an aggressive work environment has fostered a culture of interruption that must be fought.

A2.2 Exposure-Based Methods

Recognizing that maladaptive thought, emotion and behavior patterns that reach clinical significance are usually long-entrained, the focus of CBT is to have repeated new learning experiences in which the patient can practice skills acquired in therapy until they have been perfected. For example, a socially anxious patient may first practice skills during planned exposures with a warm and encouraging fellow therapist from their therapist's practice. In this low-stakes interaction, the patient can practice tolerating distress, inhibiting the urge to escape or to deploy safety behaviors, and redirecting attention outward when it turns to negative self-judgments. The patient can then generalize these skills by performing them in more difficult situations, such as with a fellow-therapist confederate who now acts impassively or hostilely, and in less controlled settings, such as with a boss or with strangers. Through repeated exposure and practice (ideally, every day), fears will typically decrease and skills will become habitual. These skills then serve as a resource the patient can draw upon if fear later returns or if a life situation elicits heightened concern about negative judgment.

In practice, to plan the exposures, first the patient and therapist will create a scale, usually from 0-10, where 0 is no distress and 10 is the most distress possible. This scale will be populated with events at different levels: perhaps for a patient who suffers from social anxiety, watching TV at home is a “0”, talking to the cashier at the supermarket is a “3”, asking a stranger for directions on the street is a “5”, making small-talk with a taxi driver is a “7”, and going to a party where they don’t know anyone is a “10”. Exposures are then planned from level 3 or so and are gradually increased: first, the patient will practice going to the supermarket and saying “hello” to the cashier each time. When this has ceased to be threatening, perhaps they will purchase a small item at a kiosk where they have to ask the attendant for the item, climbing to level “4”. After each set of exposures, the scale can be re-evaluated for all events (some might now be less threatening than they were in the past) and the patient, with the therapist’s guidance, chooses the next level/action that they find feasible to expose themselves to, and plans a new set of exposures. Depending on the diagnosis and the individual’s learning propensities (e.g., whether they learn best model-based or model-free, and what their individual learning rate is), each exposure may need to be repeated a few or more times.

A2.3 Exposures and Reinforcement Learning Prediction Errors

One reason for the gradual nature of exposures is obvious: the patient will not agree to do something too distressing, and even if they do it, they may decide to terminate the therapy due to the high level of distress. However, RL theory suggests at least two more reasons for why gradual exposure is more effective. First, rewards and punishments are not merely external: if a socially anxious patient goes to a party on their own for their first exposure, their high subjective level of distress will function as an outcome for the action. If their initial estimate of the value of the action of “going to a party” was very low, this severe-distress outcome will only confirm this prediction. As a result, there will be no prediction error, and no new learning. The exposure will have failed. Instead, a well-planned exposure at a low level of predicted distress (level 3) can lead to learning to the extent that the exposure is planned so well that the patient experiences less distress than expected. This can be achieved by discussing all possible outcomes and their cognitive appraisal, and planning for how to mitigate any distress that arises. The goal is for the patient to experience that the event that they predicted would lead to level 3 distress will not be as bad as they thought, and distress may even dissipate to level 0 over time without escaping the situation. This will cause a prediction error that will result in learning.

A second advantage of gradual exposure is that prediction errors are not too large. Research suggests that large prediction errors prompt the creation of a new state (Gershman, Blei, Niv 2010). This new state will then be updated with the new information, but the old value will remain unchanged. According to RL theory, to unlearn old maladaptive state and action values, one should experience prediction errors that are not too small (otherwise there will be no learning) and not too large (as they will cause state-splitting). CBT exposures seem tailored to deliver exactly such prediction errors.

References

Abramowitz JS: Variants of exposure and response prevention in the treatment of obsessive-compulsive disorder: A meta-analysis. Behavior Therapy 27(4):583-600, 1996

Arkadir D, Radulescu A, Raymond D, et al: DYT1 dystonia increases risk taking in humans. elife, 5, e14155, 2016

Barto AG: Adaptive critic and the basal ganglia, In: Models of Information Processing in the Basal Ganglia. Edited by Houk JC, Davis JL, Beiser DG, Cambridge, MA, MIT Press, 1995, pp. 215-232

Beck AT: Cognitive Therapy and the Emotional Disorders. New York, NY, New American Library, 1976

Beck AT: The current state of cognitive therapy: a 40-year retrospective. Archives of General Psychiatry, 62(9):953-959, 2005

Beck JS: Cognitive Behavior Therapy: Basics and Beyond. New York, NY, Guildford Press, 2011

Behrens TE, Woolrich MW, Walton ME, Rushworth MF: Learning the value of information in an uncertain world. Nature Neuroscience 10(9):1214-1221, 2007

Bennett D, Silverstein SM, Niv Y: The two cultures of computational psychiatry. JAMA Psychiatry 76(6):563-564, 2019

Brown R and Kulik J: Flashbulb memories. Cognition 5(1):73-99, 1977

Burke CJ, Tobler PN, Baddeley M, Schultz W: Neural mechanisms of observational learning. Proceedings of the National Academy of Sciences 107(32):14431-14436, 2010

Butler AC, Chapman JE, Forman EM, Beck AT: The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clinical Psychology Review 26(1):17-31, 2006

Cockburn J, Collins AG, Frank MJ: A reinforcement learning mechanism responsible for the valuation of free choice. Neuron 83(3):551-557, 2014

Cohen ZD: Treatment Selection: Understanding What Works For Whom In Mental Health. Publicly Accessible Penn Dissertations, 2932, 2018

Cohen RT and Kahana MJ: Retrieved-context theory of memory in emotional disorders. bioRxiv, 817486, 2019

Craske MG, Treanor M, Conway CC, et al: Maximizing exposure therapy: An inhibitory learning approach. Behaviour Research and Therapy 58:10-23, 2014

Daw ND: Trial by trial data analysis using computational models, in: Decision Making, Affect, and Learning: Attention and Performance XXIII. Edited by Delgado MR, Phelps EA, Robbins TW. Oxford, NY, Oxford University Press, 2011

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

Foa EB: Prolonged exposure therapy: Past, present, and future. Depression and Anxiety 28(12):1043–1047, 2011

Foa EB and Kozak MJ: Emotional processing of fear: exposure to corrective information. Psychological Bulletin 99(1):20-35, 1986

Gershman SJ, Blei DM, Niv Y: Context, Learning, and Extinction. Psychological Review 117(1):197–209, 2010

Gershman SJ, Jones CE, Norman KA, et al: Gradual extinction prevents the return of fear: implications for the discovery of state. Front Behav Neurosci 7:164, 2013

Gershman SJ, Monfils M-H, Norman KA, Niv Y: The computational nature of memory modification. eLife 6, 2017

Gershman SJ, Norman KA, Niv Y: Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences 5:43–50, 2015

Gershman SJ, Radulescu A, Norman KA, Niv Y: Statistical Computations Underlying the Dynamics of Memory Updating. PLoS Computational Biology 10:(11), e1003939, 2014

Gillan CM, Kosinski M, Whelan R, et al: Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife 5:e11305, 2016

Hare TA, O'Doherty J, Camerer CF: Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. Journal of Neuroscience 28(22):5623-5630, 2008

Hofmann SG, Asnaani A, Vonk IJ, et al: The efficacy of cognitive behavioral therapy: A review of meta-analyses. Cognitive Therapy and Research 36(5):427-440, 2012

Huys QJ, Maia TV, Frank MJ: Computational psychiatry as a bridge from neuroscience to clinical applications. Nature Neuroscience 19(3):404-413, 2016

Madan CR, Ludvig EA, Spetch ML: Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. Psychonomic Bulletin & Review 21(3):629-636, 2014

McGuire JT, Nassar MR, Gold JI, Kable JW: Functionally dissociable influences on learning rate in a dynamic environment. Neuron 84(4):870-881, 2014

Meyer V: Modification of expectations in cases with obsessional rituals. Behaviour Research and Therapy 4:273-280, 1966

Montague PR, Dayan P, Sejnowski TJ: A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience 16(5):1936-1947, 1996

Montague PR, Dolan RJ, Friston KJ, Dayan P: Computational psychiatry. Trends in Cognitive Sciences 16(1):72-80, 2012

Moutoussis M, Shahar N, Hauser TU, Dolan RJ: Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. Computational Psychiatry 2:50-73, 2017

Nassar MR, Rumsey KM, Wilson RC, et al: Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience 15(7):1040, 2012

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

Niv Y, Edlund JA, Dayan P, O'Doherty JP: Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. Journal of Neuroscience 32(2):551-562, 2012

Powers MB, Halpern JM, Ferenschak MP, et al: A meta-analytic review of prolonged exposure for posttraumatic stress disorder. Clinical Psychology Review 30(6):635-641, 2010

Reynolds JN, Hyland BI, Wickens JR: A cellular mechanism of reward-related learning. Nature 413(6851):67-70, 2001

Rouhani N and Niv Y: Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. Psychopharmacology 236(8):2425-2435, 2019

Schultz W, Dayan P, Montague PR: A neural substrate of prediction and reward. Science 275(5306):1593-1599, 1997

Shohamy D and Daw ND: Integrating memories to guide decisions. Current Opinion in Behavioral Sciences 5:85-90, 2015

Sutton RS and Barto AG: Time-derivative models of Pavlovian reinforcement, in Learning and Computational Neuroscience: Foundations of Adaptive Networks. Edited by M. Gabriel M and Moore J, Cambridge, MA, The MIT Press, 1990, pp 497-537

Sutton RS and Barto AG: (2018). Reinforcement learning: An introduction. MIT press.

Talmi D, Lohnas LJ, Daw ND: A retrieved context model of the emotional modulation of memory. Psychological Review 126(4):455, 2019

Wilson RC, Geana A, White JM, et al: Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General 143(6):2074, 2014

Wittmann BC, Daw ND, Seymour B, Dolan RJ: Striatal activity underlies novelty-based choice in humans. Neuron 58(6):967-973, 2008

Yu AJ, Dayan P: Uncertainty, neuromodulation, and attention. Neuron 46(4):681-692, 2005