Molecules and Circuits

From “Intrusive Thinking: From Molecules to Free Will,” edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

From “Intrusive Thinking: From Molecules to Free Will,” edited by Peter W. Kalivas and Martin P. Paulus.
Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

2

What Are the Circuits That Mediate and Update Intrusive Thinking?

Paul E. M. Phillips and Amy L. Milton

Abstract

This chapter discusses psychological constructs considered to be central to the mediation of intrusive thinking and the neural circuits that underlie these processes. It assimilates intrusive thoughts with conditioned responses, discerns associate structures that can support these responses, and suggests how episodic information may be integrated with these associations. Mechanisms by which intrusive thoughts can be updated are explored, with a focus on extinction and memory reconsolidation. Intrusive thoughts ultimately engage many areas of the brain as they encompass sensory, cognitive, motor, and somatic processes. In this chapter, the focus is on specific circuits within the prefrontal-limbic network that are proposed to encode, update, and maintain the content of intrusions. These circuits include interconnecting pathways between the ventral tegmental area, nucleus accumbens, medial prefrontal and orbitofrontal cortices, hippocampus, and the amygdaloid complex.

Introduction

To begin to answer the question posed in the title, we must first consider what intrusive thinking is. Specifically, we need to address the nature of the content of an intrusive thought: Is it an episodic representation? Could it be a visceral urge, an emotive state that is devoid of episodic content? There is unlikely to be a single answer, as intrusive thinking is ultimately made up of different compositions of these extremes, specific to the underlying pathology. For example, delusions in schizophrenia are clearly episodic in nature, often composed of a detailed and elaborate narrative. By contrast, in obsessive-compulsive disorder (OCD), obsessions come more in the form of an urgency state that is often likened to anxiety. While there may be episodic aspects to this process, such as obsessing over specific contexts, the debilitating qualities of

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

the obsession emerge from the anxiety-like state. Likewise, obsessive behavior in substance use disorders and other addictive disorders comes in the form of craving, an emotive state which, in its fundamental form, is devoid of episodic content. This state, however, is intimately attached to episodic representation of experiences related to the addictive behavior. Further, posttraumatic stress disorder (PTSD) is also associated with an anxiety-like state that is very clearly linked to fragmented episodic memories. This differentiation between the content of intrusive thoughts may, on the surface, seem subtle but is likely to be critical when its neural substrates are considered because the circuits that process episodic information are separable from those underlying emotive states. These differences fall into a common dichotomy in the control of behavior; namely, there are parallel systems that subserve cognitive functions. This dichotomy has a classic separation of processes that can be loosely summarized in the form of a speed–accuracy trade-off: fast, low-computational processes that amount to estimations are generated in parallel to more precise computations that have a higher cognitive demand. We and others have equated this dichotomy to parallel processes used in machine learning that are classified as model-free and model-based computations (Clark et al. 2012) as we believe this is an intuitive and tractable framework. The basis of this separation is the complexity of the information that is stored to support a learned association. A model-based computation establishes an environmental model that can be used to explore potential and inferred connections between stimuli and states. In contrast, a model-free computation is a single-dimensional value assigned to a stimulus based on the reliability of its association with a motivationally relevant outcome.

Central to this line of reasoning, intrusive thoughts are often triggered by environmental stimuli. For instance, in substance use disorders, drug craving can be elicited by drug cues and is posited to become stimulus bound during the transition to addiction (Tiffany and Carter 1998). Accordingly, drug seeking in rodent models, a proxy for craving, can be elicited by unconditioned stimuli (e.g., abused substances or stressors) or conditioned stimuli (e.g., drug cues or conditioned stressors). In PTSD, intrusive episodes are often linked to environmental stimuli in a manner consistent with overgeneralization (i.e., when otherwise neutral stimuli elicit a threat response). Hence, intrusive thoughts often take the form of Pavlovian-conditioned responses, which is our focus in this chapter. As an aside, it is worth noting that while intrusive thoughts can be triggered by unconditioned stimuli, they are unlikely to be *unconditioned responses* because they are not elicited in naïve individuals but rather develop with psychiatric pathology. Importantly, where tested, stimulus-driven intrusions exhibit similar neural signatures to those that are not triggered by an explicit external stimulus (M. C. Anderson, pers. comm.).

What neural circuits are necessary to support these associations? We will discuss circuits that support Pavlovian associations, both for emotive responses and those that incorporate representation of stimulus properties. We

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

will also consider structures that mediate the storage and retrieval of episodic memories as well as the interactions between all of these circuits. We will make the case that circuitry, including midbrain dopamine neurons that project to the nucleus accumbens in the ventral striatum (mesolimbic pathway) or to the medial prefrontal cortex (mesocortical pathway), is necessary for at least a subset of emotive associations. Thereafter, discussion of the neural circuitry will be broadened to include substrates that support Pavlovian associations that can support inferential reasoning, with a specific focus on the central role of the orbitofrontal cortex (OFC) in these higher cognitive processes. We will also explore medial temporal and frontal structures implicated in the acquisition and consolidation of episodic memories and discuss circuits involved in the process of updating existing associations, both through extinction and memory reconsolidation.

Mesocorticolimbic Dopamine- and Emotive-Conditioned Responses

Many psychiatric disorders that are associated with intrusive thinking (e.g., schizophrenia, substance use, OCD, and PTSD) have been linked to perturbations in dopamine transmission in the striatum and/or the prefrontal cortex. These clues have driven extensive research on dopamine transmission with regard to psychiatric disorders, including those where intrusive thinking is a prominent feature.

Reward Prediction Errors

In the mid- to late 1990s, computational neuroscientists and computer scientists who shared an interest in learning algorithms came to the hypothesis that dopamine transmission provides a critical teaching signal for stimulus–reward associations in a model-free learning algorithm (Barto 1995; Montague et al. 1996). This notion was most famously linked to the empirical research of Wolfram Schultz et al. (1997). While this area has expanded enormously since then, the computational role of dopamine transmission in reward learning is most commonly ascribed to variants of the temporal difference algorithm (Sutton and Barto 1998). This algorithm is a time-derivative model that evolved from simple trial-by-trial learning models developed to account for animal behavior. The hypothesized role of dopamine neurons in the model is to signal discrepancies between reward expectation and rewards received to update the current expectation of reward: firing increases when rewards are larger than expected and decreases when rewards are smaller than expected. This model puts dopamine in a critical role in the acquisition and maintenance of stimulus–reward associations. In addition to this putative role in learning, mesolimbic dopamine transmission has consistently been shown to invigorate

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

the enactment of conditioned responses (Flagel et al. 2011b; Ostlund and Maidment 2012).

Model-Free and Model-Based Processes

In a simple cue–reward learning task, where there is spatial separation of the cue and the reward, some animals will approach the cue during its presentation ("sign tracking") whereas others will approach the place where the reward will subsequently be delivered ("goal tracking"). This latter conditioned response requires cognitive representation of the spatial location of future reward, consistent with a model-based process. Interestingly, when animals emit these model-based conditioned responses, dopamine release in the nucleus accumbens does not follow the canonical prediction error signal (Flagel et al. 2011b). This suggests that the specific pattern of signaling is selective for model-free computations. Furthermore, while intact dopamine transmission is necessary to perform either the conditioned response or to acquire sign-tracking conditioned responses, it is not necessary to acquire goal-tracking responses. Similarly, disrupting dopamine transmission during the Pavlovian-to-instrumental transfer disrupts general transfer (invigoration) but it does not affect the cue's ability to bias action selection specifically toward the cue-associated reward, which requires a model-based representation. Recent experiments, however, have revealed a role for dopamine-encoded prediction error signals in some model-based processes (Sharpe et al. 2017). Specifically, it was demonstrated that phasic dopamine signals participate in the updating of the stimulus–stimulus association that takes place in the absence of motivationally valent stimuli and can be used to generate model-based inferences (sensory preconditioning). Thus, mesolimbic dopamine transmission has a somewhat nuanced role in Pavlovian processes, with some predilection for simple stimulus–reward associations. It appears to be important in the acquisition and updating of model-free stimulus–reward associations as well as some, but not all, model-based associations. In addition, mesolimbic dopamine has universal psychomotor-activating properties by which it invigorates the response to conditioned stimuli; however, through this, it is thought to convey only the emotive properties of the association rather than specific (sensory) properties of the conditioned or associated unconditioned stimulus.

Aversive Signaling

To date, the focus has been on associations with appetitive stimuli. Needless to say, intrusive thoughts are often evoked by aversive stimuli. The role of the mesolimbic and mesocortical dopamine systems in computations relating to aversive information has been much more controversial. In many cases, it is simply assumed that aversive stimuli will be computed in this learning model in a similar manner to stimuli that predict lower than previously

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

expected rewards. However, the evidence for this type of encoding is mixed. Mirenowicz and Schultz (1996) observed minimal responses to mildly aversive stimuli (air puffs to the hand) in midbrain dopamine neurons of nonhuman primates. In contrast, Matsumoto and Hikosaka (2009) observed robust changes in the activity of dopamine cells on the presentation of aversive stimuli. In this latter work, the investigators reported some populations of dopamine neurons that encoded aversive information by changing their firing rates in the opposite direction to reward information. Specifically, the presentation of unexpected aversive stimuli or conditioned stimuli that increased the expectation of aversion resulted in reduced firing of these neurons. However, Matsumoto and Hikosaka also reported populations of putative dopamine neurons that increased their firing rate to predictions of aversion. These neurons tended to reside in dorsolateral aspects of the dopaminergic ventral midbrain (dorsolateral substantia nigra pars compacta). They responded similarly to stimuli that increased the expectation of aversion to those that increased the expectation of reward, an encoding pattern often referred to as an unsigned prediction error. However, this coding pattern is not without controversy. Specifically, Fiorillo (2013) has argued that the observed positive responses to aversive stimuli relate to their sensory properties rather than their motivational salience. Others have also questioned whether all of the recorded neurons in these studies are truly dopamine-containing neurons. To address this concern of neuronal-type specificity, Cohen et al. (2012) recorded the firing rates of genetically identified dopamine neurons in mice in response to presentations of appetitive- and aversive-related stimuli. They reported that modulation of the firing rates to aversive-related stimuli were consistently in the opposite direction to those for reward stimuli. These neurons, however, were exclusively recorded in the ventral tegmental area (i.e., the ventral medial aspect of the ventral midbrain) and so did not include the homologous anatomical region from which the unsigned prediction error signals were reported by the Hikosaka group. In neurochemical studies that measure dopamine levels in terminals, results with aversive stimuli have also been mixed. Studies measuring dopamine levels over minutes tend to report increases in dopamine to the presentation of aversive stimuli, especially in the prefrontal cortex (Young 2004; Butts et al. 2011). In contrast, measurements on the order of seconds reveal decreases in dopamine in the nucleus accumbens to aversive stimuli (Roitman et al. 2008). Some investigators addressing this issue have argued that increases in dopamine transmission following an aversive stimulus observed over minutes were responses to the relief from aversion at the offset of the stimulus rather than a response to the onset (Ungless 2004).

Integration of Appetitive and Aversive Information

With systematic interrogation of dopamine neurons in the ventral tegmental area based on identified afferent and efferent connectivity, Lammel et al. (2012)

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

proposed that appetitive and aversive processing by dopamine neurons is segregated into subcircuits. They demonstrated that activation of a circuit connecting the lateral habenula, ventral tegmental area, and the medial prefrontal cortex produces an aversive state, while activation of another circuit connecting the laterodorsal tegmentum, ventral tegmental area, and lateral shell of the nucleus accumbens produces an appetitive state. Complementary to these data, Tye and colleagues recently reported that a set of prefrontal-projecting dopamine neurons were selectively activated by aversive stimuli, whereas dopamine neurons that projected to the nucleus accumbens were activated by rewards (Vander Weele et al. 2018). While this schema does not account for all of the apparent discrepancies in the literature, it does move the field forward toward a resolution. However, the issue of whether and how dopamine transmission integrates appetitive and aversive information in a manner that could instantiate a single, unitary learning model is still an open question. Given our current understanding, this integration could be through one of several possibilities, including bidirectionality of appetitive and aversive information by mesolimbic dopamine neurons (Roitman et al. 2008; Cohen et al. 2012), gleaning appetitive and aversive information from different populations of dopamine neurons (Lammel et al. 2012; Vander Weele et al. 2018), or combining appetitive information from dopamine neurons with aversive information from other neural substrates (Daw et al. 2002).

Broader Circuity

When considering the entirety of the pathways which support these stimulus–stimulus associations, the complexity of the circuitry rapidly expands. In addition to the laterodorsal tegmentum and the lateral habenula, the central nucleus of the amygdala, in particular, has been implicated as important upstream circuitry to dopamine neurons in model-free processes (Clark et al. 2012). Nonetheless, it is important not to dismiss the rich convergent inputs into the ventral midbrain from many areas of the brain (Geisler and Wise 2008). Indeed, an optimal temporal difference algorithm should have access to all available sources of predictive information about rewards and punishers as well as information on motivational states.

More Complex Stimulus–Stimulus Associations

Orbitofrontal Cortex

From a plethora of research, it is evident that the OFC encodes many different features of motivational stimuli (Thorpe et al. 1983; Padoa-Schioppa and Assad 2006; Stalnaker et al. 2014). Indeed, it seems that just about any aspect of perception is encoded in about twenty percent of OFC neurons! This

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

multidimensional encoding seems to be especially important to support model-based associations that permit inferential reasoning. For example, OFC lesions do not affect the acquisition of simple Pavlovian associations or the ability for animals to avoid food that has been paired with illness or fed to satiety. However, unlike intact controls, OFC-lesioned animals continue to respond to stimuli that predict these devalued outcomes (Gallagher et al. 1999). These reinforcer devaluation studies demonstrate that the OFC contributes to the animal's ability to derive the updated expected value of a cue by linking the previously learned stimulus–reward association with the current incentive value of that outcome without having yet directly experienced the pairing of this cue with the devalued outcome. Consistent with the OFC playing a selective role for model-based computations, OFC lesions do not affect general Pavlovian-to-instrumental transfer, but they do disrupt the cue's ability to selectively enhance responding for the specific outcomes the cues predict (Ostlund and Balleine 2007a). While the computational role of the OFC in these processes is not yet precisely defined, a reasonable inference would be that the OFC is used to evaluate the content of model-based associations upon their retrieval rather than being directly involved in the acquisition or storage of this information per se. In this regard the OFC and dopamine may have parallel, complementary roles in response to conditioned stimuli with dopamine conveying emotive responses and OFC relaying stimulus-specific sensory information.

Episodic Information

While the above discussion differentiates associations that represent specific features of conditioned or unconditioned stimuli, these more complex associations do not necessarily incorporate episodic information. However, since intrusive thoughts are often episodic in nature, it is pertinent to expand our consideration to circuits involved in the processing of episodic information. The study of episodic memory is a particularly rich area of neuroscience and has primarily focused on the medial temporal lobe, including hippocampus formation as well as the prefrontal cortex and to some extent the frontal and parietal cortices. However, investigations of the interactions between Pavlovian associations and episodic memory is relatively sparse. Nonetheless, a particularly elegant series of experiments from Shohamy and colleagues have provided evidence that the hippocampus has central roles in associative processes. For example, they showed that hippocampal episodic information can provide a framework for model-based processes to permit inferential reasoning (Wimmer and Shohamy 2012). They also demonstrated that the hippocampus can drive dynamic corticostriatal network connectivity governing stimulus–reward associations (Gerraty et al. 2018). These innovative studies have started to build a foundation for understanding the use and integration of episodic information into learned associations. This platform could be particularly useful for studies of intrusive thinking.

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

Updating and Inhibition of Intrusive Thought

Extinction

To think about the updating of intrusive thoughts, we should consider the different plasticity processes that can be engaged following the consolidation of an intrusive thought or memory. Broadly speaking, the neural trace or "ensemble" that encodes a thought or memory can be triggered to induce retrieval, reconsolidation, or extinction of the trace under similar, but importantly different, conditions. Extinction has been the most extensively studied, having been originally described by Pavlov (1927). Extinction is operationally defined as the degradation of behavior that was previously supported by a learned association. It takes place when the reliability of the association is weakened, typically occurring after extensive exposure to the unreinforced cue. Extinction can be modeled as a reduction in the strength of an association between a condition and unconditioned stimulus in a simple bidirectional learning system, such as the model-free system putatively associated with dopamine transmission. However, for many associations, it has been argued that extinction learning is not simply the "unlearning" of an association but new, discriminative learning that associates different states of the world with new contingencies in a more complex model. While there are clear demonstrations that extinction *can* be new learning—most likely dependent on prefrontal cortical regions—it perhaps should not be assumed that this is a universal mechanism. At least at the level of the amygdala, molecules usually associated with the depotentiation of synapses increase in activity during Pavlovian extinction, and these molecules are also necessary for extinction to occur (Merlo et al. 2014). Thus, it remains a possibility that different associations have fundamentally different mechanisms of extinction. Regardless of mechanism, a defining feature of intrusive thinking is that the underlying associations are relatively resistant to extinction.

Reconsolidation

The resistance to extinction of intrusive memories may be especially maladaptive if, instead of engaging extinction, reactivation of the intrusive thought leads instead to strengthening of ensemble via reconsolidation mechanisms. Reconsolidation is a process, more recently described than extinction, that can also potentially act to update memories. Reconsolidation refers to the induction of memory lability (or, more mechanistically, memory "destabilization") under certain conditions of retrieval, and the subsequent "restabilization" of the trace in a strengthened or updated form. The restabilization phase is dependent upon protein synthesis in a similar manner to the original consolidation process when the memory was first stored. Therefore, preventing reconsolidation following memory retrieval (e.g., by administration of a protein synthesis

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

inhibitor) can essentially erase the memory. Reconsolidation has been investigated in the context of fear memories (Nader et al. 2000), where it was shown to be dependent on protein synthesis in the basolateral amygdala. Likewise, the basolateral amygdala is necessary for the reconsolidation of drug-related associations that support place conditioning and conditioned reinforcement (Milton et al. 2008a, b; Théberge et al. 2010). Interestingly, protein synthesis in the nucleus accumbens core is also required for reconsolidation of associations supporting conditioned drug place preference but is not required to reconsolidate drug-cue associations that support conditioned reinforcement (Théberge et al. 2010). The extent to which the disruption of reconsolidation in one node of a motivational network can impinge on the function of other nodes in that network is a question that has been surprisingly understudied. Despite this, it has been hypothesized that reconsolidation could provide a mechanism by which memories can be strengthened (Lee 2008), generalized (Vanvossen et al. 2017), and integrated into wider memory networks (Hardt et al. 2010). Therefore, if reconsolidation mechanisms were engaged upon reactivation of an intrusive thought, it is possible that a form of mnemonic "positive feedback" could be established, by which a reactivated intrusive thought not only becomes strengthened when it restabilizes, but generalizes and integrates with other associative traces to lead to an increase in the number of cues and contexts that could trigger the intrusive thought. The consequent increase in the likelihood of triggering the thought would potentially increase the likelihood of subsequent reactivation, leading to further strengthening, generalization, integration, and so on.

Therapeutic Strategies Using Updating Mechanisms

In addition to providing a potential pathological mechanism underlying the persistent and recurrent nature of intrusive thought, reconsolidation could also provide a therapeutic target. While the administration of protein synthesis inhibitors to humans to disrupt reconsolidation is not straightforward, it is possible to capitalize on the updating function of memory reconsolidation in designing therapeutic approaches. For example, developing an approach first used in the preclinical literature (Monfils et al. 2009), Schiller et al. (2010) used a "retrieval-extinction" procedure in which they reactivated a fear memory and subsequently extinguished this memory while it was destabilized (in the "reconsolidation window"). Similarly, James et al. (2015) used visuospatial interference (playing the video game *Tetris*) to disrupt the reconsolidation of intrusive mental images produced by watching traumatic films clips. Although further research needs to be conducted to determine the mechanisms by which these procedures produce effects, including corroboration at the molecular level (Cahill and Milton 2019), these approaches hold potential for the development of new treatment for neuropsychiatric conditions characterized by intrusive thoughts. As a cautionary note, an important consideration for these

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1

types of approaches is that different types of stimulus–stimulus associations (i.e., emotive and cognitive) could exist in parallel, supported by independent neural circuits. Therefore, which memory or, more specifically, which aspect of the association drives the pathology may be critical. With some pathologies, for instance, extinguishing an intrusive episodic memory could be futile if the untreated emotive association reattaches to a new episodic memory.

Conclusions

The neural circuits that mediate and update intrusive thoughts are complex but potentially tractable based on our current understanding of model-based and model-free systems and their operation. It is important to appreciate, however, that these circuits are not fixed and immutable, but rather it is likely that they undergo repeated rounds of plasticity and metaplasticity, leading to imbalance within the circuit. In this way, we hypothesize that an adaptive physiological process, supported by functional neural circuitry, can become persistent, recurrent, and pathological.

Acknowledgments

This work was supported by NIH grants R01-DA039687, U01-AA024599, and P50-MH106428-5877 to Paul Phillips, and U.K. Medical Research Council Programme Grant MR/N02530X/1 to Amy Milton. Amy Milton is further supported by the Ferreras-Willetts Fellowship in Neuroscience at Downing College, Cambridge.

From "Intrusive Thinking: From Molecules to Free Will," edited by Peter W. Kalivas and Martin P. Paulus. Strüngmann Forum Reports, vol. 30, Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-54237-1