# 13

# Systems Approach to Intrusive Experiences

Angela C. Roberts, Rita Z. Goldstein, David Badre,
Bernard W. Balleine, Hugo D. Critchley,
Aikaterini Fotopoulou, Sophia Frangou,
Karl J. Friston, Tiago V. Maia, and Elliot A. Stein

## Introduction

This chapter explores how intrusive experiences may occur at a systems level from psychological, computational, neurobiological, and physiological perspectives. A general scheme is proposed of the essential elements of an intrusive experience, and where in this scheme dysregulation could occur to increase the likelihood of an intrusive experience. It also considers a range of psychological and mathematical models that have been applied to explain how intrusions may ultimately happen, some of which are more closely integrated into neurobiological systems than others. These include a Bayesian model of active inference, integrated psychological and physiological models of interoception, and psychological and neurobiological models of working memory and associative learning and their relevance to concepts of flexibility and stability.

## Phenomenology of Intrusive Experiences

Human mental operations (e.g., perception, emotion, cognition, metacognition, and action planning) are both complex and diverse. It is therefore important that we clearly define the phenomenological properties of intrusive thinking, particularly because it can encompass a wide array of forms, topics, and themes (see Visser et al., this volume). Here, we employ the term *intrusive experience* instead of intrusive thinking to denote that our deliberations apply to intrusive verbal thoughts, intrusive nonverbal thoughts (e.g., images,

**Group photos (top left to bottom right)** Angela Roberts, Rita Goldstein, Bernard Balleine, Sophia Frangou, Hugo Critchley, Elliot Stein, Tiago Maia, Katerina Fotopoulou, Karl Friston, Elliot Stein, Bernard Balleine, Rita Goldstein, Hugo Critchley, Karl Friston, David Badre, Katerina Fotopoulou, Angela Roberts, Tiago Maia, Sophia Frangou, David Badre, Angela Roberts

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

music), intrusive impulses (e.g., motor actions), as well as intrusive bodily sensations.

Intrusive experiences have been conceptualized in varying ways (Rachman and Hodgson 1980; Parkinson and Rachman 1981; Salkovskis and Harrison 1984; Edwards and Dickerson 1987b; Freeston et al. 1991; Yao et al. 1999), and an overall consensus definition is currently lacking. The most common features across definitions involve the involuntary and disruptive nature and internal attribution of intrusive experiences; with regard to valence and controllability, there is greater variation (Rachman and Hodgson 1980; Parkinson and Rachman 1981; Salkovskis and Harrison 1984; Edwards and Dickerson 1987b; Moulding et al. 2014). Rather than attempting to provide a general definition of intrusive experiences—a goal that has tended to elude the field and has been tackled in more detail by Visser et al. (this volume)—we focus on three stages inherent to intrusive experiences (Figure 13.1). This deconstruction allows for the empirical probing of the processes and neural systems that underlie intrusive experiences, with the ultimate goal of identifying the most appropriate targets for intervention when such experiences become pathological. Consequent upon this model are the following parameters:

- The intrusion itself is inherently neutral. It is conceptualized here as a neural event (or cascade of events), the origins of which are likely to be relatively localized within specific brain circuits or networks.
- Intrusions undergo appraisal. During appraisal, attributes are assigned to the intrusion. By definition, intrusions are unintended and thus they will be appraised as involuntary. Assignment of other attributes and emotional responses to the intrusion will depend on its nature, content, and context (situational and personal) in which it occurs.
- Post-appraisal cognitive control mechanisms (Braver 2012) determine the response strategy to the intrusion.
- The resulting intrusive experience is not inherently pathological but rather a common universal human experience (Salkovskis and Harrison 1984; Freeston et al. 1991; Corcoran and Woody 2008; Bouvard et al. 2017). Some intrusive experiences, however, can be pathological, depending on their nature, content attributes, recurrence, controllability, and behavioral consequences (e.g., Julien et al. 2007; May et al. 2015).

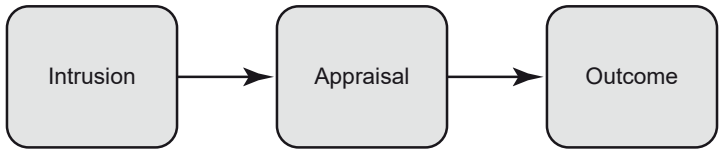



**Figure 13.1** Components of the intrusion experience.

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 Intrusion**

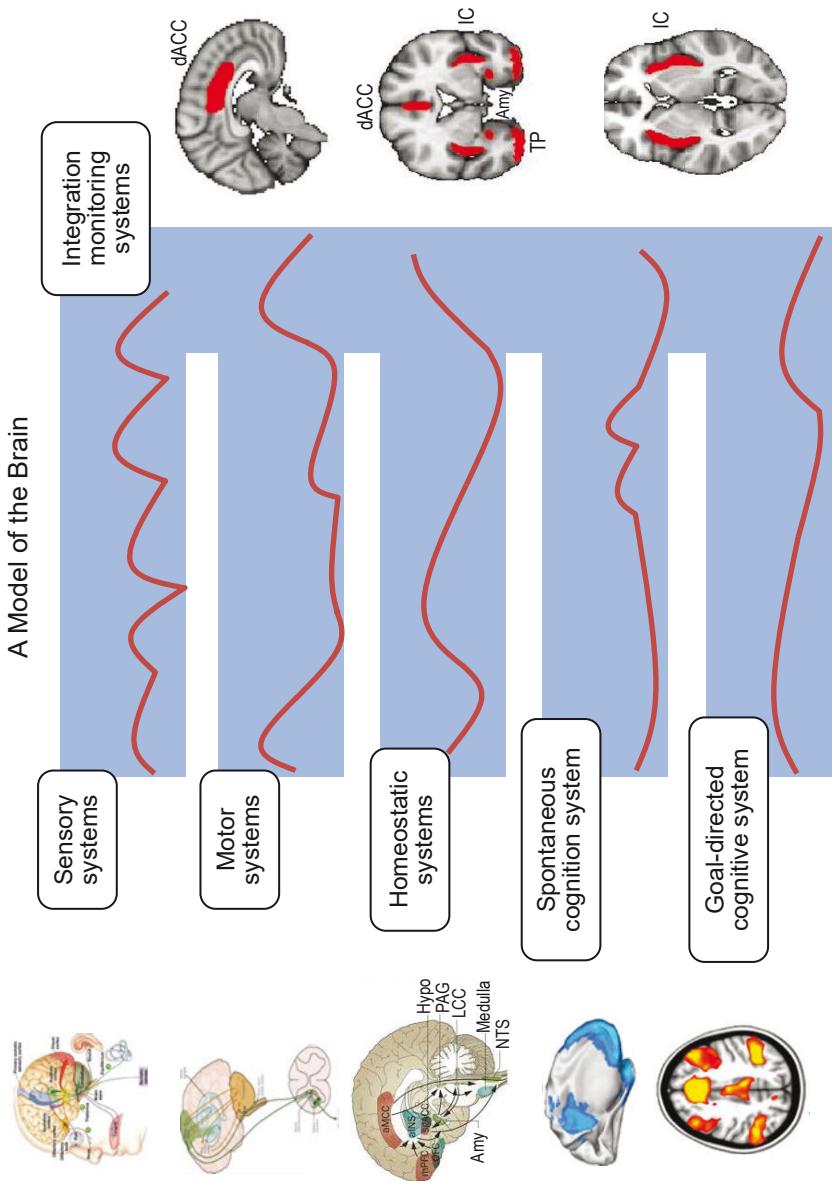
An intrusive experience has a neural "locus of origin," is sufficiently strong so that it spreads to brain regions with which it is closely linked, and propagates beyond a critical threshold which allows it to interrupt other processes and enter awareness.

The locus of origin can provide an intuitive account for the nature and content of the intrusive experience (Figure 13.2). Intrusive experiences of a sensory nature (e.g., images, music) are likely to originate within sensory systems. Intrusive experiences that involve movement are likely to originate in motor systems. Intrusive experiences that involve somatic sensations (e.g., thirst) are likely to originate in homeostatic systems (Figure 13.2; for a discussion of different neurological intrusion domains, see Gourley et al., this volume).

Tourette syndrome is a good example of where a locus of origin for the intrusive experiences can be identified. In Tourette syndrome, intrusive premonitory sensations and movements (i.e., tics) are associated with abnormal activation in somatosensory and motor cortical regions (Conceição et al. 2017). A locus of origin formulation is more challenging for intrusive experiences involving verbal thoughts. Recent advances in cognitive neuroscience, however, suggest that cognition in everyday life is dominated by thoughts that are not directly linked to sensory processing or task-directed behavior (Kane et al. 2007). Several terms (e.g., spontaneous cognition, unconstrained cognition, or mind wandering) are currently used to refer to these stimulus- and task-independent processes. In parallel, emerging neuroimaging findings have associated spontaneous cognition with connectivity within the default mode network, a functional brain network that is more active during stimulus- and task-independent periods (Andrews-Hanna et al. 2010; Dixon et al. 2014). However, it is important to note that our model postulates that regardless of the initial locus, the originating signals spread to additional brain regions following connectivity pathways so that intrusive experiences acquire multisystem associations once they reach a certain threshold (see below). In other words, they enter the global workspace (Dehaene et al. 1998) or form part of the winning coalition (Maia and Cleeremans 2005).

To account for how intrusive experiences occur, we propose two heuristic mechanisms: a breach and a permissive mechanism (Figure 13.3). These mechanisms are described separately although they may coexist. They are conceptually embedded within theories that view experience as the outcome of selective signal propagation in the face of competition (Dehaene and Changeux 2004; Beck and Kastner 2009; Graziano and Webb 2015) or global constraint satisfaction (Maia and Cleeremans 2005). The mechanisms for signal selection are currently unclear and have been described with various terms, including signal biasing or weighting (Sergent and Dehaene 2004; Beck and Kastner 2009), signal enhancement (Graziano and Webb 2015), biased competition (Desimone 1998; Deco and Rolls 2005), and gating (as 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




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

**Figure 13.2** The brain is functionally organized into cognitive systems supported by spatially defined networks (Power et al. 2011). This simplified model illustrates brain activity within brain systems (left) that could host a potential "locus of origin" of an intrusion: sensory and motor systems, homeostatic systems involved in the integration of exteroceptive and interoceptive signals (Salvato et al. 2020), a system involved in spontaneous cognition supported by the default mode network (Raichle et al. 2001), and a system for goal-directed behavior supported by frontoparietal regions (Fox et al. 2005). The model posits that intrusions are experienced as such when they enter "awareness," which is likely to occur in the presence of additional recruitment of networks involved in monitoring (shown on the right), such as the salience network (Seeley et al. 2007). The blue areas (middle) represent the network space, with oscillatory activity therein denoted by red lines. Amy (amygdala), dACC (dorsal anterior cingulate cortex), Hypo (hypothalamus), IC (insula cortex), PAG (periaqueductal gray), LCC (lateral cerebral cortex), NTS (nucleus tractus solitarius), TP (temporal pole).

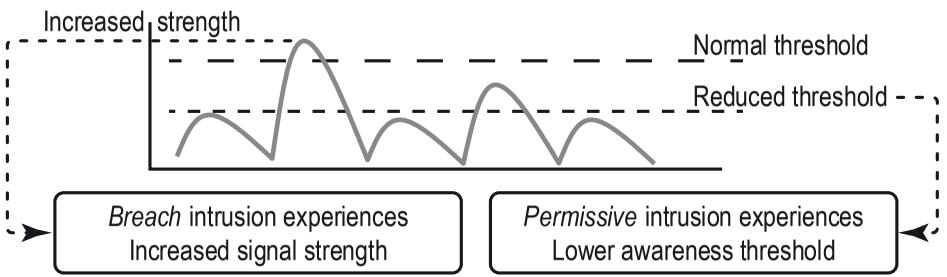



**Figure 13.3** Breach and permissive intrusive experiences.



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

discuss in more detail below). Here we use the analogy of *awareness threshold* (borrowed from sensory perception) to visualize the moment a signal gains sufficient *biological momentum* to breach the threshold of awareness. Accordingly, breach intrusive experiences can occur because the strength, features, or contextual significance of the originating signal enables its selective enhancement. In contrast, permissive intrusive experiences occur when the threshold is transiently or persistently lowered and thus permits the propagation of weaker signals. The timing of intrusive experiences (i.e., when they occur) can be influenced at any point by the external environment as well as by internal states, which can be referred to as "motivational states" in that they combine representations of somatic states and overall general behavioral drives (discussed in more detail below). It also follows that intrusive experiences are influenced by genetic and molecular factors, including neurotransmitters (e.g., Bonvicini et al. 2016; Sinopoli et al. 2017), that define healthy within-individual variation (i.e., the likelihood of an intrusion within an individual) and interindividual differences (i.e., differences between individuals in the likelihood of experiencing intrusions), and that these may be associated with pathological conditions affecting brain integrity at multiple organizational levels (e.g., Keelan et al. 2019).

Generally, signals relating to survival (e.g., hypoglycemia) will generate breach intrusive experiences. The same could apply to abnormally generated signals, as in the case of Tourette syndrome, where abnormal sensorimotor activation spreads to other brain regions (e.g., the insula) and eventually breaches the threshold of awareness (Conceição et al. 2017). Signals relating to significant prior (e.g., childhood abuse, traumatic event) or immediate circumstances (e.g., negative thoughts about the self) may also be selectively enhanced and thus breach the awareness threshold. In such cases, the content of the intrusive experiences is more likely to be "personal" to the individual. The personal nature of the intrusion is also likely to constrain the range of its content; thus, such intrusive experiences are likely to be stereotypical. The intrusion experiences observed in posttraumatic stress disorder (PTSD) are prime examples as their content is repetitious and of personal significance (American Psychiatric Association 2013). Our model also predicts that permissive intrusive experiences are likely to have a more variable and circumstantial content because the lowering of the awareness threshold will permit the propagation of a variety of signals. Attention deficit hyperactivity disorder (ADHD) would be a prototypical example of a condition in which permissive intrusive events might occur. Currently, intrusive experiences in ADHD are considered in terms of abnormalities in attentional brain systems that gate awareness (Castellanos and Proal 2012; Bozhilova et al. 2018). As already mentioned, the dichotomization of intrusive experiences as breach or permissive does not imply that they are mutually exclusive. For example, up to 50% of patients with Tourette syndrome have ADHD, suggesting that breach and permissive intrusions may co-occur and determine clinical severity and complexity.

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 Appraisal**

During the appraisal stage, we postulate that the intrusive experience will attract unconditional and conditional attributes and emotional states. By definition, intrusive experiences will be unconditionally labeled as involuntary as they bypass processes of agency (see Liu and Lau, this volume; Gallagher 2012; Moore and Fletcher 2012; Braun et al. 2018). However, typical intrusive experiences retain the "sense of ownership"; that is, the sense of selfhood we attribute to our own bodily sensations, thoughts, and actions (Gallagher 2012). It is worth noting that they are distinct from psychotic experiences which, although often construed as intrusive (especially hallucinations and delusions), typically involve a loss of agency and self-ownership (Feinberg 1978; Moore and Fletcher 2012; Frith 2014).

The appraisal of intrusive experiences is a multisystem phenomenon that may, in some cases, rely on complex representations involving semantic/linguistic networks. During appraisal, the attributes assigned to intrusive experiences and the emotional responses they invoke will depend on their content, nature, and normative significance (i.e., alignment of personal beliefs and societal values) (Korsgaard 2009). We argue that the ultimate purpose of the appraisal is to determine the "likedness" of the intrusive experience; that is, the degree to which the experience is aligned with the individual's future plans (Figure 13.4). As used here, likedness aligns with notions of motivational relevance (Higgins 2011) and self-congruence (Rogers 1959; Higgins 1987) and, as mentioned above, the appraisal of the intrusive experience depends on the characteristics of the individual having the experience, including their exposures.

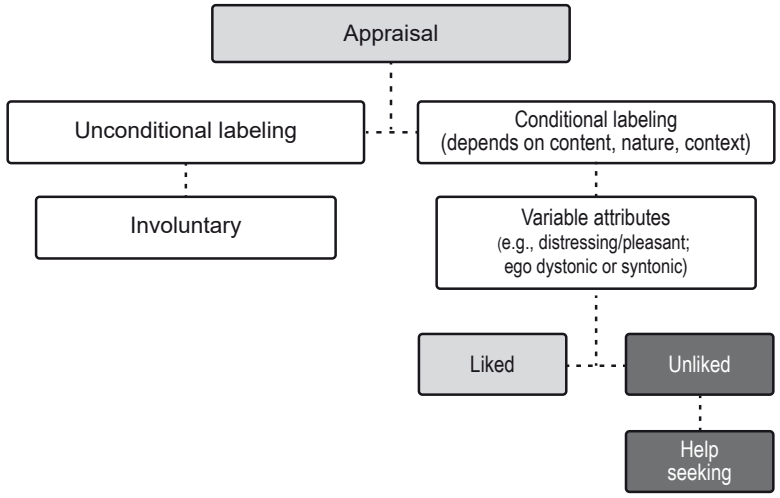



**Figure 13.4** Appraisal of intrusive experiences.

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

Intrusive experiences that are appraised as distressing and "not liked" are more likely to be classified as clinically significant and to elicit help-seeking behavior. However, intrusive experiences can be of a positive nature and liked, as in thoughts associated with loved ones or that emerge from sudden insight or "eureka" moments (Kounios and Beeman 2014). Still, intrusive experiences that are deemed positive are not always adaptive and may contribute to further pathology by providing confirmation for maladaptive beliefs, as in hedonic hunger in individuals with restrictive eating disorders (Lowe et al. 2016).

## The Outcome

The outcome of the appraisal will invoke mechanisms and networks that support selective attention, decision making, response inhibition, and response selection (Niendam et al. 2012; Langner and Eickhoff 2013; Zhang et al. 2017; Chen et al. 2018b). We assume that there will be no voluntary inhibition for liked intrusive experiences (Figure 13.5). The experience would either be allowed to decay or it could be maintained through attentional mechanisms. A liked intrusive experience may even act as a catalyst or starting point for another mental or motor plan. In such cases, the switch from the pre-intrusion state to a new one may be viewed as a positive outcome of the intrusive events. Eureka moments would fall under this category.

By contrast, "unliked" intrusive experiences will evoke attempts at voluntary inhibition. The success or failure of the experience will depend on the functional integrity of frontostriatal networks that are generally implicated

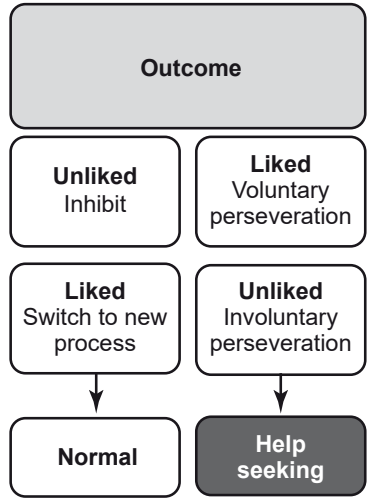



**Figure 13.5** Outcome of intrusive experiences.

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

in inhibitory control (Niendam et al. 2012; see also chapters by Balleine and Badre, this volume; Bari and Robbins 2013). Of note, disorders characterized by intrusive experiences also present with a more general impairment of inhibitory control that affects multiple aspects of cognition and behavior (Gourley et al., this volume; Marsh et al. 2009; Shin et al. 2014; Morand-Beaulieu et al. 2017; Pievsky and McGrath 2018). Failure of inhibitory control is expected to give rise to perseveration and/or premature action (as exemplified in compulsivity and impulsivity, respectively), which may elicit secondary appraisals involving frustration, anger, and increased arousal directed at the failure to inhibit rather than the original intrusive experience. Such an outcome is likely to increase the allocation of attentional resources to the intrusive experience and the inhibitory failure; in some individuals, this may reinforce intrusion experiences, leading to a pathological loop.

## Salience, Precision, and Value in Intrusive Experiences

Having considered the nature of intrusive experience in terms of definitions, phenomenology, and their implications in a clinical setting, this section provides a complementary perspective that takes its lead from systems neuroscience and, in particular, computational approaches that offer a formal and quantitative account of the phenomenology at hand. We introduce concepts of *precision*, *salience*, and (motivational) *value* that may help understand how and why intrusive experiences occur. We use two working examples that illustrate how dysregulation within these psychological domains may explain different sorts of intrusive experiences; namely, those associated with obsessive-compulsive disorder (OCD) and PTSD. This section concludes with a discussion of the conceptual implications in terms of computational architectures that underwrite intrusive experience, and how accompanying computational models and (neuronal) process theories can be used to characterize empirically observed behavioral and neuronal responses.

### Brief Review of Active Inference with a Special Focus on the Nature of Precision, Salience, and Value

The treatment in this section considers cognition as a process of inference or belief updating in the brain. Specifically, we use an active inference framework to cast action and perception as solving an inference problem; namely, optimizing beliefs about states of affairs in the lived world and, crucially, beliefs about how the world should be sampled or navigated (i.e., beliefs about plans or actions). In short, we make a simplifying assumption that trains of thought can be associated with planning as inference (Attias 2003; Baker et al. 2009; Botvinick and Toussaint 2012; Baker and Tenenbaum 2014; Mirza et al. 2016).

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

Planning as inference rests on an internal model of how (unobserved) states of affairs causing (observed) sensations are generated. This is known as a *generative model*, usually expressed mathematically in terms of the *likelihood* of some observations, given latent or *hidden states* and *prior beliefs* about those states (for a more detailed account, see Appendix 13.1 and Figure 13.A1). In this setting, beliefs are nonpropositional (i.e., subpersonal) and simply refer to probability distributions encoded by synaptic activity or connectivity (in the sense of Bayesian belief updating or belief propagation). A simple example of active inference is the way that we forage for visual information. If I move my eyes from one position to another, the state of my oculomotor system will change, and this will have profound implications for the sensory impressions on my retina. A sequence of eye movements would then correspond to a particular *policy* or action strategy. My job is to infer the most likely policy that "something or someone like me" would engage, and then select a particular action (i.e., a next move) under that policy.

In selecting the most likely policy, I will necessarily refer to my prior beliefs about the policies I am likely to pursue; namely, those that provide the most evidence for my (generative) model of the world. This can be expressed formally in terms of a prior over policies, based on *expected free energy*. Free energy, in this instance, is known as an *evidence bound* in machine learning (Winn and Bishop 2005) and can be thought of as a measure of expected surprise or prediction error. Mathematically, expected surprise is also known as uncertainty. This means that I will select those policies (and implicit courses of action) that resolve uncertainty about the state of the world. This formulation of active inference emphasizes the two-way exchange between an agent and her world, where the implicit action-perception cycle means effectively that beliefs can change states of the world, which in turn change the sensations that update beliefs. For an illustration of this circular causality, see Figure 13.A2 in Appendix 13.1.

### *Motivational Value and Salience*

Mathematically, expected free energy can be decomposed in a number of ways (see Figure 13.A2 for a decomposition into *risk* and *ambiguity*). For our purposes, the more prescient decomposition is in terms of *salience* and *value*. Heuristically, the (negative) expected free energy of a policy is equal to salience plus value (see Appendix 13.1 for details and how this decomposition relates to other disciplines in neuroscience):

$$\text{Expected free energy} = \text{Salience} + \text{Expected value}. \tag{13.1}$$

In this setting, *salience* corresponds to the uncertainty resolving or intrinsic (epistemic) value of a policy. It is variously referred to as relative entropy, mutual information, information gain, Bayesian surprise, intrinsic motivation,

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

or value of information (Barlow 1961; Howard 1966; Optican and Richmond 1987; Linsker 1990; Itti and Baldi 2009). Salience, therefore, reflects the information gain or resolution of uncertainty afforded by response to a cue: "How much will I learn, if I look over there?"

Salience can be contrasted with expected (extrinsic or instrumental) value, which is the motivational value of a policy defined in terms of outcomes that are preferred a priori. Expected value is an important construct in optimal control theory in engineering, reinforcement learning in psychology, and utility theory in economics. Expected value simply scores the expected returns (cf. rewards) following a particular policy, expressed in terms of the log probability of some prior preferences (i.e., preferred or expected outcomes). This is sometimes referred to as extrinsic value, as opposed to epistemic value, to make it clear that these are extrinsically supplied outcomes that provide a motivational value for the policies under consideration. The foregoing offers a definition of salience and motivational value in terms of active inference and the accompanying quantities or functionals of Bayesian beliefs encoded by neuronal activity and connectivity. So, what about *precision*?

### *Precision and Attention*

Precision is an attribute of sensory outcomes or evidence at hand. Very precise data are informative, in the sense of having a high signal to noise. The important thing, from our perspective, is that precision has to be estimated or inferred in a context-sensitive fashion. For example, if I know that I am exploring an unfamiliar room in the dark, I know that the precision of visual sensations will be much lower than the precision of my somatosensory sensations. I would therefore assign a greater precision to the mapping between the hidden states of the world (e.g., "a chair in front of me") and the somatosensory outcomes (e.g., "I will feel this chair if palpated"). Conversely, if I know the light is on, I will adjust the precision of my visual mapping such that visual information is afforded much more precision and has a much greater influence on belief updating about the state of my room.

Psychologically, this is effectively the same as attention (Desimone et al. 1990; Desimone 1998; Womelsdorf et al. 2007; Parr and Friston 2019); in other words, a selective gating or attentional filtering affords one sort of sensory stream with more precision than another. When this deployment of attention is part of a policy (i.e., a covert action much like the premotor theory of attention), we have an attentional policy that is implemented through a selective gating of the sensory information at hand (Limanowski and Friston 2018). This will become an important concept later in our discussion of memory (see below), where policies that selectively afford precision to different sources of information correspond to *gating policies*. In hierarchical generative models, the level of the implicit gating or precision control may determine the nature of attention: exogenous versus endogenous

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

(e.g., spatially directed attention vs. attention to a particular visual feature, respectively).

## Applying Computational Models of Active Inference to Understand Intrusions in Obsessive-Compulsive Disorder and Posttraumatic Stress Disorder

According to the above formulation, precision is a ubiquitous attribute of the likelihood and prior beliefs that nuance or select the right kind of information for belief updating. This selection or gating rests on the excitability or lateral inhibition among competing representations at any level of a hierarchical generative model. As noted above, precision is context sensitive and must be inferred; this means that it depends on beliefs about hidden states (i.e., context) and, indeed, beliefs about policies (i.e., "what I am doing"). In contrast, salience and value are attributes of a particular plan or policy whose evaluation involves belief updates about the succession of states in the future, under a particular course of action. A salient act is one that resolves uncertainty or is likely to have the greatest epistemic affordance. The value of a plan is scored by the degree to which the outcomes are likely to be realized.

Let us now consider the computational pathology that might underwrite a typical intrusive experience in OCD: "checking behavior." Assume that there are two states of the world with which I am concerned: "the door is locked" versus "the door is unlocked." Any policies that resolve uncertainty about whether the door is locked will have a high salience. If I do not know a priori whether the door is locked or not, checking whether the door is locked has the greater salience and will, in the healthy course of things, resolve my uncertainty.

Imagine now that my generative model also predicts a state of physiological arousal due to the possibility that the door is unlocked, and all the catastrophic consequences that such a state of affairs could entail. If I can resolve my uncertainty and be 100% certain that the door is locked, then I predict that the associated interoceptive evidence for physiological arousal will also be attenuated. Now, imagine what would happen if I were unable to attenuate the precision of interoceptive signals:[1] I would check the door, expecting to find it locked and *expecting my arousal to subside*, but it does not.

I would now be in the curious situation of still being uncertain about when the door is locked because I have sensory (interoceptive) evidence at hand that I cannot have checked the door (because I am still physiologically aroused). This means that the epistemic affordance of door checking is still in play. In fact, unless I can attenuate my interoceptive signals, this uncertainty will

[1] In active inference, a failure to attenuate the precision of proprioceptive or interoceptive signals is accompanied by a failure to engage motor or autonomic reflexes. In this example, a failure to engage autonomic reflexes means that a state of physiological (sympathetic) arousal would persist.

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

continue to be in play and induce successive checking behavior that may proceed indefinitely.

Notice in this example that checking behavior has been formulated in terms of aberrant salience because the action of rechecking the door does not lead to the resolution of uncertainty. This aberrant salience is suboptimal (i.e., pathological) because of a failure to attenuate or change interoceptive signals. In short, a failure of sensory attenuation led to aberrant salience and a persistent epistemic affordance that never resolves itself. In other words, no matter how many times I check the door, I never sense that my uncertainty has been resolved, which could further maintain a state of autonomic arousal. Expressed even more simply, this checking behavior is futile because there is an irreducible uncertainty about the state of the world due to a failure to attenuate interoceptive evidence from my body. This predictive processing, or active inference account of OCD, is based (and elaborates) on work by Kiverstein et al. (2019) and Rae et al. (2019a), and owes much to seminal accounts of why patients with OCD appear to be "stuck in a loop."[2] For example, Roger Pitman (1987:336) suggested that "the core problem in OCD is the persistence of high error signals, or mismatch, that cannot be reduced to zero through behavioral output," and that "the obsessive-compulsive's internal comparator mechanism is faulty. No matter what perceptual input it receives, it continues to register mismatch….It may be that in fact the action was well done, but the defective comparator cannot register it" (Pitman 1987:340).

In turn, Szechtman and Woody (2004:111) suggest that "the symptoms of obsessive-compulsive disorder...have what might be termed an epistemic origin—that is, they stem from an inability to generate the normal 'feeling of knowing' that would otherwise signal task completion." On the empirical side, Gentsch et al. (2012:656) found decreased sensory attenuation in OCD, which was suggested to "explain the tendency of individuals with OCD to continuously register error signals, and to experience dissatisfaction in outcome processing."

The somewhat contrived formulation of OCD, in terms of aberrant salience, focused on an account of intrusive experience that manifests in overt motor behavior. Does this explanation hold for intrusive thoughts, images, and experiences in PTSD? A plausible account could proceed along the following lines: Imagine that, at the point a traumatic event is experienced, there is some particular configuration of (interoceptive or exteroceptive) sensory inputs in play. The traumatic event can then induce a one-shot learning of the concomitant gating policy. When this pattern of sensations is encountered subsequently, it is extremely difficult to ignore, because sensory information is afforded great precision. These sensory cues will induce belief updating and the selection of

[2] This account is from the PhD thesis by Itzchak (Isaac) Fradkin: "Deficits in processing of prediction errors in obsessive compulsive disorder: Effects on action, thoughts, learning and agency," Hebrew University of Jerusalem, June 2019.

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 traumatic narrative or policy that entails overt or covert action. In the latter setting, action is neither motoric (i.e., mediated by striated muscles) nor autonomic (mediated by smooth muscles) but *attentional* in nature. In other words, the gating policy is called up in an obligatory fashion, sometimes described in terms of modulating *sensory* and *prior* precision (Skewes et al. 2014; Ainley et al. 2016; Powers et al. 2017; Rae et al. 2019a).

This traumatic active inference or learning will induce a recapitulation of the internal policy or narrative that may enable posterior expectations all the way down to the sensory levels of perceptual hierarchies. In other words, a triggering event will *breach* attentional thresholds and induce a cascade of hierarchical and sequential processing that recapitulates the sequential narrative associated with the original trauma. The mechanisms behind such fictive (intrusive) experience are part and parcel of self-evidencing under a generative model. Common examples here include dreaming, imagination, and the generative or constructive perceptual processing associated with structure learning and eureka moments (Hinton et al. 1995; Botvinick et al. 2009; Gershman and Niv 2010; Tervo et al. 2016; Friston et al. 2017; Gershman 2017).

Based on this account, the intrusive experience induces a gating policy that prescribes covert (mental) actions that are manifest as internal scene construction and accompanying narratives (Peters et al. 2017; Wilkinson et al. 2017), as opposed to the mostly overt actions considered in the OCD example above. Clearly, the foregoing account does not offer a qualitative distinction between intrusive experiences that reflect an adaptive response to trauma and the psychopathology that results when intrusions are experienced (or manifest) as maladaptive and persistent. However, the computational account narrows down the field, in terms of where aberrant inference and learning may be operating in conditions like OCD and PTSD. Next, we consider the failure of sensory attenuation and subsequent failure to relearn the right sort of attentional response as a plausible candidate.

## Summary

The two working examples of OCD and PTSD were introduced here to make a key point: the intrusive experience of OCD rests upon *aberrant salience* that is secondary to a *failure of sensory attenuation*; namely, an aberrant top-down modulation of sensory mappings. In contrast, the PTSD example appeals only to aberrant precision via a *breach* of sensory attenuation due to traumatic learning of a particular attentional set or gating policy. In other words, people with PTSD may lose the capacity to ignore the irrelevant and be plagued by breaches of attentional filtering or gating endowed by sensory attenuation. If one subscribes to these accounts, the conclusion is that the primary pathophysiology behind both kinds of intrusive experience is a failure of sensory attenuation that most likely involves interoceptive signals. Interestingly, a failure of sensory attenuation emerges in computational treatments of other psychiatric

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

conditions (Skewes et al. 2014; Ainley et al. 2016; Powers et al. 2017; Rae et al. 2019a); in particular, schizophrenia and autism. [For a review of aberrant precision and sensory attenuation in psychiatry, see Stephan et al. (2016) and Friston (2017) for details and references.]

On this account, a minimal but sufficient explanation for intrusive experience is a failure of inhibitory control inherent in the sensory attenuation. The key thing that the active inference framework brings to the table is that this inhibitory control is not about the contents of perceptual experience, but the precision or attention afforded this content. From a physiological perspective, this is important because a failure of inhibition (i.e., a failure of sensory attenuation or attenuation of sensory precision) may be mediated not by hyper- (or de-) polarizing neuronal populations but by modulating their excitability or gain. In turn, this suggests the mechanisms that underwrite the pathophysiology of intrusive experiences are located either in classical modulatory neurotransmitter systems or the downstream effects on cortical excitability (as mediated by fast-spiking inhibitory interneuron coupling with pyramidal cells).

In summary, the emerging picture is of a deficit in the neuromodulatory mechanisms (and dynamics) that implement the top-down control of attention; namely, its sensory attenuation. A natural corollary is that there may be as many different forms of intrusive pathologies as there are neuromodulation mechanisms and projections. Irrespective of this diversity, and the accompanying regional specificity of evidence accumulation schemes in the brain, one underlying mechanism becomes apparent: the breach of sensory attenuation (i.e., attentional filtering) by exogenously or endogenously generated cues that underwrite belief updating about states of the world and our active engagement with that world. Clearly, in many instances, this intrusion is part of normal perceptual synthesis and subsequent planning. For example, a loud noise is salient because it offers a person the opportunity to "look over there" and resolve any uncertainty associated with the surprising sensory signal.

The pathology implicit in the examples above rests on aberrant salience that maintains irreducible uncertainty incurred through a failure to attenuate interoceptive signals (as in the case of overt compulsive behavior in OCD). It can also rest on the failure of sensory attenuation to be attributed to, and subsequent failure to relearn, the right kind of attentional response to triggers (as in the case of PTSD). As discussed above, the notion of a breach in sensory attenuation is a key aspect of higher-order models of intrusive experiences that consider the evaluation (i.e., the appraisal) of inferred states and subsequent metacognitive influences. At present, three conclusions follow from the formal analysis of this section that are remarkably consistent with the treatments offered in other chapters in this volume:

- Intrusive experiences are inherently *interruptive* in the sense that they induce a quantitative change in the selection of narratives or sequential policies, which underwrite overt or covert (mental) action.

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

- These intrusive episodes (events) are *experienced* in virtue of being manifest in terms of beliefs about (overt or covert) action. This follows because there is an egocentric aspect to action generated by these beliefs; that is, the only thing that can act is "me."
- Finally, intrusive experiences have, at some level, a *salience*, either an irreducible epistemic affordance that cannot be dispelled or in terms of aberrant precision; namely, the failure to suspend or attenuate attention to certain kinds of cues. In OCD, for example, the inability to attenuate arousal sensations manifests in the repetition of salient acts (such as checking), despite the fact that these acts do not produce a lasting reduction in uncertainty about the state of the world. (A computational model of the neuronal underpinnings of recurrent intrusions in OCD is provided in Appendix 13.1.)

## The Insula and Functional Anatomy of Salience and Value

Now let us consider the above account from a systems neuroscience perspective. In this setting, salience can be thought of as an attribute of a cue (i.e., internal or external stimulus) deemed important to the individual *in a given context* (Uddin 2014; Kahnt and Tobler 2017; Miyata 2019)—it is salient because of the potential for information gain and thus belief updating. Salience is distinct from value in that the latter is a valenced or signed currency that varies monotonically from negative to positive, whereas the former is an unsigned currency (i.e., something is salient or not). This means that value and salience are dissociable in terms of what they mean for behavior: both negative and positive outcomes can be salient in the sense that experiences can change our beliefs, even if they are unpleasant (Kahnt and Tobler 2017). As such, intrusive experiences can be thought of as arising from an aberrant processing of internal and external stimuli with respect to the current (belief) state of the individual. This salience misattribution leads to an overemphasis of one thought or action over the current, ongoing cognitive process and subsequently influences attentional capture, motivation, and goal-directed cognition. Importantly, the unsigned nature of salience calculations necessitates that both appetitive and aversive stimuli can sway the calculations of salience that ultimately influence behavior.

There are many potential points at which biases can enter salience calculation. Ascribing salience to a given stimulus at a given time scale (Kennerley et al. 2011) and within a given context (Heilbronner and Hayden 2016) results from integration across a wide range of processes, including attentional (Menon and Uddin 2010), reward (Olney et al. 2018), affective (Etkin et al. 2011), and homeostatic regulation (Craig 2009).

Neurobiological instantiation of both ongoing and intrusive, highly salient events occurs at many levels of the neuraxis. One highly interconnected hub that seems to play a major role as an integrator or transmitter of the interoceptive and

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

exteroceptive environment is the insula. More specifically, the anterior insula (and the von Economo neurons it contains) possesses the anatomical linkages to support awareness and (together with its connections to, e.g., the anterior cingulate; Craig 2009) the monitoring of the environment that is necessary (when combined with the calculation of value) to assign behavioral relevance to the event. In Tourette syndrome, for example, the insula may play a role in assigning salience and aversiveness to premonitory urges (Conceição et al. 2017).

There are important, mostly bidirectional connections between the anterior insula and key affective, cognitive, autonomic, and regulatory systems—components which place the anterior insula in a unique position in the calculation of salience (Critchley et al. 2005; Craig 2010; Nieuwenhuys 2012). In addition to the posterior regions of the insula that receive predominantly somatosensory inputs, homeostatic regulators enter via the hypothalamus and amygdala, hedonic inputs from the nucleus accumbens and orbitofrontal cortex, and motivational, social, and cognitive information from anterior cingulate, ventromedial, and dorsolateral prefrontal cortex (Craig 2010). While anatomically and functionally simplistic, this schema provides a framework uniquely placing the insula in the position to assess the relative weights of the environmental processes in the assessment of attentional capture. The insula also has important connections to motor regions that allow it to then drive behavior. In fact, in Tourette syndrome, it may play a role in driving tics (Conceição et al. 2017), whereas in addiction it may be driving craving (Naqvi and Bechara 2010; Naqvi et al. 2014).

Importantly, there appears to be a transdiagnostic component to the dysregulation of salience attribution (McTeague et al. 2016), as neuroimaging studies have demonstrated anterior insula involvement across a number of diagnostic assignments in various disorders characterized by intrusive events, including addiction (Naqvi et al. 2014), ADHD (Klein et al. 2013; Bubenzer-Busch et al. 2016; Norman et al. 2016), autism (Gu et al. 2018), OCD (Zhu et al. 2016), psychosis (Brosey and Woodward 2017), anxiety (Paulus and Stein 2006; Shiba et al. 2017), and depression (Ellard et al. 2018). This is potentially important because it suggests that interoception plays a key role in all forms of aberrant salience or precision attribution, as illustrated by the OCD example above, and will be discussed in more detail below.

## Interoceptive Contributions to Intrusive Experiences

Understanding intrusive experience at the level of brain systems will be incomplete without a consideration of the systems that underlie the self. Selfhood is fundamental to the phenomenology of intrusive experiences (see Liu and Lau, this volume). If intrusive experiences are to be understood as involuntary mental phenomena that disrupt ongoing psychological narrative flow (see above), one needs to have a sense of oneself as both an observer, experiencing such

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

intrusions, and as an agent perceiving the intrusions as unsolicited interruptions of one's sense of agency (e.g., I did not intend to have these thoughts, or perform these actions, even though I recognize them as my thoughts, or my actions). Importantly, the brain systems that regulate and represent bodily physiology, or interoception, are considered to be at the core of selfhood.

Increasingly, there is appreciation that the self as a continuous coherent, unitary representation is not the output of one specific single specialized system within the brain. Experiences of selfhood are embodied and require coordination between dissociable brain systems, as revealed by careful experimentation and in the symptomatic expression of particular psychiatric and neurological "disorders of self," such as depersonalization disorder (for a review, see Fletcher and Fotopoulou 2015). Within a broad taxonomy, selfhood can be parsed into minimal (embodied or biological) and extended (reflective and narrative) components (akin to the first- and higher-order mechanisms described by Liu and Lau, this volume). The sense of a core minimal self is proposed to emerge from the integrative processing of sensory and motor signals from the body. The frequent concomitant occurrence of sensory signals on the body eventually gives rise to mental, predictive models of "owned" first-person feelings of (bodily) sentience and presence (e.g., I exist and feel alive in this body), agency (e.g., I was the author of this action), and ownership (e.g., this bodily experience belongs to me) (Gallagher 2005; Seth et al. 2012). Extended concepts of the self are built on embodied self-representation to encompass the notion of the narrative or autobiographical self (Damasio 1999). Extended selfhood affords the ability to make one's self the object of explicit thoughts irrespective of any particular experience or perspective in the here and now (e.g., self-reference, I am a woman). More generally, this enables reflection on one's experiences across time, space, and person in counterfactual ways: I have always been a woman, I anticipate being a woman tomorrow, I imagine that I am a woman in the mind of others (Fotopoulou 2015). These notions rest on the idea that while perception can be understood as the unconscious process of hierarchical Bayesian inference on the (hidden) causes of sensory input, more higher-order abilities for self (metacognition) or other mentalization or reflection rest upon similar unconscious inferential processes of greater depth, whereby the generative models refer not only to current sensory predictions but also to predictions about the effects of actions, not yet executed, and bodily or external situations, not yet encountered (e.g., Palmer et al. 2015).

Interoception is at the core of this hierarchical view of the self and, by extension, of psychopathological disorders of self-representation (for an overview, see Khalsa et al. 2018). Indeed, there is growing theoretical acknowledgment that core aspects of selfhood (Seth 2013; Fotopoulou and Tsakiris 2017) and emotion (Gu et al. 2013; Seth 2013; Barrett et al. 2016) can be formalized as the inferential processing of interoceptive signals, implemented within a Bayesian/predictive coding framework (discussed above). Correspondingly, models of how interoceptive inference is instantiated or regulated within the

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

brain can inform understanding of emotional and psychosomatic disorders, such as anxiety, depression, and fatigue (Paulus and Stein 2006; Barrett et al. 2016; Stephan et al. 2016). Here we offer an overview of the neurocognitive mechanisms by which interoceptive processes underpin self-representation, psychopathology, and, in particular, intrusive experiences. We conclude with an example of an eating disorder as an instance in which interoception influences the psychopathological expression of intrusive experiences within a hierarchical predictive framework that encompasses self-conceptualization.

## What Is Interoception?

Interoception encompasses afferent signaling, integrative processing, and central representation of the internal physiological state of the body (Quadt et al. 2018; for a discussion of alternative definitions, see Ceunen et al. 2016). Interoception is the sensory component of homeostatic and allostatic control. Homeostasis refers to the regulation of internal physiology, through which life is sustained by maintenance of a more or less constant internal environment, through supporting the dynamic metabolic needs of bodily tissues while excluding potential toxic or other threats to the integrity of the body (homeostatic regulation; Cannon 1929). Cardiac output, blood oxygenation, hydration, temperature regulation, and blood glucose are among the many parameters regulated homeostatically. However, homeostatic interoceptive autonomic reflex arcs alone are inefficient: better control of internal state is achieved through allostasis, wherein the future state of the body is predicted and responses are made in anticipation of future physiological states to mitigate unpredicted dyshomeostatic states that threaten life (Sterling 2012).

Allostasis is informed by the integration of interoceptive information with exteroceptive (about the external world) information for the predictive selection of autonomic/physiological and behavioral action or "policies," which ultimately ensure longer-term survival. For example, the set point of the homeostatic baroreflex, which stabilizes blood perfusion of organs by regulating the heart's beat-to-beat output, is allostatically adjusted to meet actual and anticipated physical demands (e.g., if you see a snake or bear in the woods, baroreflex suppression allows your heart rate and blood pressure to rise together to enhance skeletomuscular perfusion, facilitating the capacity for fight and flight). From a more computational perspective, the most efficient way to regulate homeostatic risk is to build a model of the body as separate from its external environment, following the cybernetic idea that "every good regulator of a system must be a model of that system" (Conant and Ashby 1970). Ultimately, physiological control combines allostatic and homeostatic mechanisms, but both can be subsumed under homeostasis (Ramsay and Woods 2014). Allostatic anticipatory control requires an inferential model (hypotheses about the causes of interoceptive inputs) of our own current and future (counterfactual) bodily states in relation to states of the external world (including

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

other agents). The complexity is reduced by holding a set of prior "beliefs" or more broadly generative models. Deviations from homeostatic ranges are avoided by choosing in advance an appropriate sequence of actions ("policies"). These can be autonomic as well as behavioral and can cross different systems. For example, you need to eat before you faint and you need to store fat for future metabolic needs when resources may need to be allocated to other tasks. These ideas are coherent with formal frameworks of brain function, such as the Bayesian brain and active inference (discussed above). Details and discussion of the neural organization supporting interoceptive processing can be found in Appendix 13.1.

## Interoception and Intrusive Experiences

There are at least three ways in which interoception can impact upon intrusive experiences:

1. It can provide *context* which can (a) have an impact on the permissive threshold for the occurrence of intrusions (discussed above) and (b) influence or constrain the content of what intrudes.
2. It can affect appraisal and control processes engaged by the intrusive experience.
3. It can also act as content itself.

Moreover, these can interact to produce a self-sustained cycle of intrusive experiences. In conceptualizing the impact of interoception on intrusive experiences, it is helpful to conceptualize it within a hierarchical or dimensional framework (see Table 13.1). Lowest in the hierarchy are the levels of physiological arousal (indexed by heart rate, blood pressure, or electrodermal activity) and the bodily changes governed by homeostatic reflex arcs. These signal the integrity and arousal state of the body through visceral afferent pathways. Fluctuations in central signaling of bodily physiology (including both engagement of ascending neuromodulatory systems and representation within primary "viscerosensory" insula, a cortical level) can thus provide the context (Pt. 1 from the above list).

As a context, psychophysiological states (e.g., sickness, arousal, and alertness) gate what enters the sensorium (Pt. 1a). For example, a heightened state of cardiovascular arousal enhances the detection and appraisal of threat (Garfinkel et al. 2014; Pezzulo et al. 2018) associated with symptoms of anxiety; increased sympathetic electrodermal tone enhances occurrence of tics in Tourette syndrome (Nagai et al. 2009). In addition, however, a particular homeostatic context, such as hunger, can motivate relevant intrusions about food (Pt. 1b; a specific example is given in the next section). Affective state represents a more elaborated interoceptive context that can again change the permissible threshold of intrusion.

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

**Table 13.1** Dimensions of interoceptive measurement, adapted after Garfinkel et al. (2015). Psychological dimensions of interoception are given in boldface; self-referential dimensions are capitalized.

| Dimensional level | Nature | Index measures |
|---|---|---|
| **EXECUTIVE** | Behavioral | Shifting from interoceptive to exteroceptive attention (e.g., within dual tasks or between tasks) |
| **METACOGNITIVE** | Correspondence between subjective self-report and objective performance accuracy | Receiver operating characteristic curves between task performance and rated confidence<br>Correlational measures of task and confidence scores<br>Trait measures (e.g., correspondence between task performance and body perception questionnaire score) |
| **SENSIBILITY** | Subjective self-report | Confidence measures on interoceptive tasks<br>Questionnaires probing interoceptive sensitivity |
| **Accuracy** | Objective behavioral performance score | Heartbeat detection tasks<br>Respiratory resistance load detection<br>Water load task<br>Balloon dilation of stomach/colon |
| Preconscious impact on other processes | Behavioral, neural | Cardiac modulation of eyeblink startle<br>Cardiac modulation of fear<br>Respiratory modulation of memory |
| Afferent signal | Neural | Visceral afferent nerve recording<br>Intracranial recording<br>Heartbeat evoked potential<br>Respiratory evoked potential<br>Neuroimaging |
| Bodily response | Organ-level response | Heart rate, heart rate variability, tachygastria, blood pressure, glucose, $O_2$ and $CO_2$ levels, etc.<br>Autonomic psychophysiology |

Second, appraisal and control processes engaged by the intrusive experience are impacted by higher-order cognitive levels of interoceptive representation, likely supported within insula and cingulate cortices (Pt. 2). Higher-order cognitive or psychological levels of interoception (highlighted in bold in Table 13.1) refer to attention and appraisal directed at bodily processes themselves. These encompass measures of interoceptive accuracy of objective (behavioral) sensitivity to bodily responses, subjective (i.e., self-reported) interoceptive sensibility to bodily signals, and metacognitive interoceptive insight (Garfinkel et al. 2015). The latter two align with notions of expectation and interoceptive prediction error ("surprise") and the precision weighting of interoceptive inputs, beliefs, and policies. Interoceptive self-efficacy (Stephan et al. 2016) is a metacognitive representation of self-efficacy.

Third, such a mental representation of bodily sensation may act as the content of the intrusion (Pt. 3). Salient bodily signals (e.g., breathlessness, heart arrhythmia, urge to void, or visceral pain) necessarily attract attention and appraisal. Upon appraisal, prior experience will determine if the intrusion per se represents a major concern or acts as a driver for subsequent general perseverative intrusions associated with overall health (e.g., health anxiety). Related to this are the so-called quasi-interoceptive signals, such as rib pain (a somatic sensation), which can be misinterpreted as a prelude to a heart attack, with anxiety again becoming amplified by the accompanying interoceptive sensations of cardiorespiratory arousal as a consequence of the appraisal process (Clark et al. 1997). Moreover, ephemeral interoceptive sensations can (through prior associations) trigger emotional (e.g., panic or fear response PTSD) or drug-related intrusive experiences such as craving (Goldstein et al. 2009; Garavan 2010). Similarly, the interoceptive feelings of premonitory urge, linked again to representation within insular cortex, will trigger tics in Tourette syndrome (Rae et al. 2018, 2019b).

Finally, an executive dimension of interoception contributes to intrusions mostly through appraisal control processes (Pt. 2) which in turn can affect the stickiness of the context (Pt. 1) and the interoceptive content (Pt. 3). The executive dimension encompasses the capacity to shift between interoceptive representations or away from interoceptive representations, aligned with both precision weighting and policy selection. Such a capacity may be evident in measures of lower levels of interoceptive signaling: for instance, heart rate variability (a product of baroreflex regulation) is linked to more general psychophysiological flexibility and is positively associated with success in suppressing unwanted intrusive thoughts and memories, like in PTSD (Gillie and Thayer 2014; Gillie et al. 2015). Conversely, intrusive perseverative cognition (worries and ruminations) and the capacity for thought control are coupled to the inflexibility associated with blunted heart rate variability, both in wakefulness and during sleep (Brosschot et al. 2010; Meeten et al. 2016; Ottaviani et al. 2016; Ottaviani et al. 2017). It should also be noted that in addition to all of these direct and indirect effects 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

interoception on intrusive experiences, interoceptive signals may be evoked under some circumstances as countermeasures to control intrusion, often through physiological relaxation but sometimes using physiological arousal (Nagai 2015).

Given this intimate relationship between interoception and intrusive experiences, it is perhaps not surprising that disordered interoceptive processing is reported across conditions associated with intrusive thinking. In anxiety disorders, increasing evidence indicates an association between anxiety symptoms and a mismatch between subjective (sensibility) and objective (accuracy) measures of cardiac interoception—a metacognitive interoceptive deficit (trait interoceptive prediction error) (Garfinkel et al. 2015) that is also relevant to symptoms in Tourette syndrome (Rae et al. 2019b), autism (Garfinkel et al. 2016), and, if extended to measures of choice, addiction (Moeller et al. 2014). Moreover, intrusive dissociative experiences, consistent with a fundamental self-disturbance in self-representation, are associated with lower-level interoceptive abnormalities (Schulz et al. 2016). Below we present an example of how abnormalities in interoception can act as the content and character of an intrusive experience.

## Intrusive Experiences in an Ego-Syntonic Disorder Exemplified by Anorexia Nervosa

Patients with anorexia nervosa report thoughts, bodily experiences, and mental images that they consider as involuntary and intrusive to other goals, even though these may not always be unpleasant in themselves and may, in fact, constitute most people's everyday experiences. For example, a patient described the feeling of a full stomach as intruding on her mental concentration (Skårderud 2007:127):

> Some days ago, I should have had a meeting with my boss. I was anxious about this. Then I decided to vomit. I couldn't stand having the lunch in my stomach. I cannot have anything in my stomach, because then I cannot concentrate. I need to be empty to feel alert.

Similar experiences of hunger or satiation and other interoceptive sensations are frequently experienced as intrusive by individuals with anorexia, while their attempts to control their eating and body weight and to "silence" any relevant bodily needs are seen as compatible with the goal of building a coherent and stable self. This treatment-resistant concordance in eating disorders between symptoms and a sense of self is referred to as ego-syntonicity (Gregertsen et al. 2017). Unlike in (ego-dystonic) disorders like OCD, where symptoms are seen as intruding into one's other everyday goals, anorexia exemplifies a psychiatric disorder where symptoms are not viewed as intrusions into one's life; instead, necessary bodily functions, and particularly interoceptive experiences, are experienced as intrusive.

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

Drawing on the framework outlined earlier in the chapter as well as from knowledge of the brain systems that support homeostatic and allostatic control (outlined above), intrusive experiences in eating disorders, particularly in anorexia nervosa, can be understood as a failure to model interoceptive states at a homeostatic level, due to a deeper failure in the regulation of metabolism (notably adiposity or fat storage) at an allostatic level (i.e., a failure to optimize flexibly the precision weighting of allostatic control policies). In other words, patients with anorexia nervosa may not be able to correctly predict and regulate adiposity (and metabolism more generally), leading to a chronic dyshomeostatic state that evokes aberrant metacognitive beliefs about the low efficacy of their autobiographical self: "I cannot eat now because I will then lose control over my eating and store excessive fat" (see Figure 13.6). Recent converging evidence highlights wide dysregulation across neuromodulatory systems in eating disorders, including hormones and neuropeptides involved in the regulation of metabolic states (see Figure 13.6; Gorwood et al. 2016), and a large-scale genetic study implicating metabolic (alongside psychiatric) factors in pathoetiology of anorexia nervosa (Watson et al. 2019). Neurocomputational formulations of allostasis, that is, predictive, counterfactual interoceptive control (Stephan et al. 2016), suggest that allostasis requires a temporary change or suspension of homeostatic set points, effectively altering the priors (beliefs) of the relevant homeostatic reflex arc (e.g., the expectation of a meal will drop

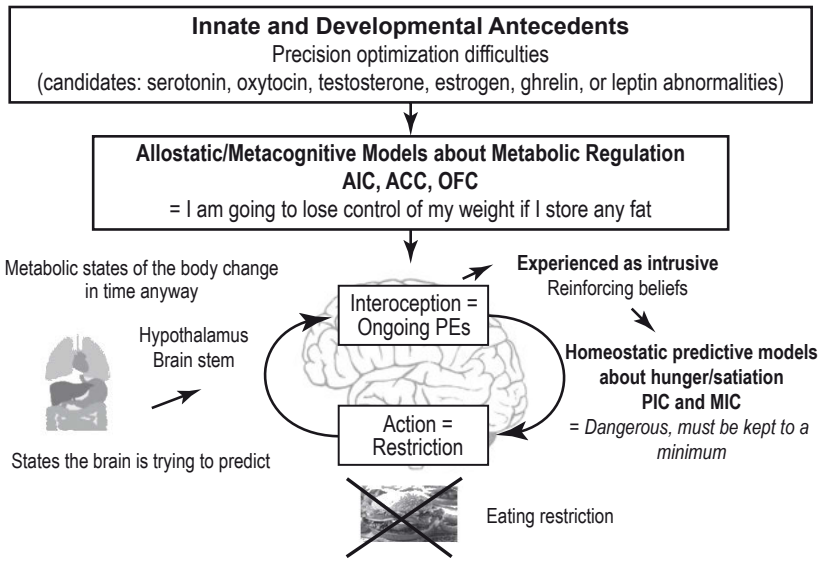



**Figure 13.6** Schematic depiction of a predictive coding account of intrusive interoceptive experiences in anorexia nervosa: AIC (anterior insular cortex), ACC (anterior cingulate cortex), OFC (orbitofrontal cortex), PIC (posterior insulate cortex), MIC (medial insular cortex), PE (prediction error).

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

blood glucose levels to mitigate the hyperglycemia that follows eating). In the brain, allostatic coupling of behavioral policy with internal physiology is supported by regions including the anterior insula and dorsal and subgenual anterior cingulate cortices. These manifest three properties (Stephan et al. 2016): (a) access to estimates of bodily state (interoception), (b) the capacity to generate predictions over longer time scales, and (c) anatomical connections (descending visceromotor outputs) that can convey sustained changes in homeostatic beliefs instantiated by more reactive humoral/autonomic reflex response arcs within hypothalamus, brain stem, and periphery. Thus, functional abnormalities within these regions may lead to inappropriate adjustments to specific physiological parameters (e.g., glucose levels before or after meals), leading to persistent prediction errors driving abnormal eating habits. For anorexia nervosa, eating control will always be suboptimal for regulating metabolic and interoceptive states since these necessarily fluctuate in time. Persistent exacerbated interoceptive feelings of hunger and satiation are experienced as ongoing intrusive experiences that interfere with the ego-syntonic goal of a rigid control of body fat, achieved by eating restraint, exercise, and/or vomiting. These acts in themselves and their interoceptive consequences reinforce the homeostatic beliefs of patients regarding the unpredictable and intrusive nature of hunger and satiation signals.

Several studies have indeed shown abnormalities in correctly predicting and experiencing interoceptive states in anorexia nervosa, including, for example, cardiac signals, satiation and affective touch, and the related brain function abnormalities best tracked by the anterior insular cortex and related limbic and prefrontal areas (Crucianelli et al. 2016; Bischoff-Grethe et al. 2018; Khalsa et al. 2018). Such abnormalities have been linked to persistent prediction errors about interoception and a dysregulated ability to adequately sense what is happening in the body resulting in a turbulent reference state; that is, a "noisy baseline" (Paulus and Stein 2010). This may explain why patients experience all those states as *intrusive experiences of the body* that need to be controlled by eating restriction, exercise, or vomiting (see Figure 13.6). These attempts to actively restrict and control hunger and satiation in turn lead to starvation and further maintenance mechanisms (starvation dampens hunger and slows down cognitive processing along with further complications). According to the above speculations, a fundamental difficulty in reducing interoceptive uncertainty via allostatic control would be at the heart of why otherwise normal feelings of hunger or satiation are experienced as intrusive and as "out of control."

## Relevance of Stability and Flexibility to Intrusion Experiences

Balancing stability and flexibility in the brain is critical for individuals to maximize exploitation and exploration of their environment. Working memory and

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

associative learning models provide a psychological and neural framework in which the concepts of flexibility and stability can be understood (Hochreiter and Schmidhuber 1997; Frank et al. 2001; Oberauer 2013). Biophysically detailed computational models have also investigated how dynamical interactions between different neuronal populations in cortex may promote stability versus flexibility (Durstewitz et al. 2000; Wang 2001). In addition, a large body of empirical evidence has implicated specific neural structures and neuromodulators in behavioral flexibility (Robbins 2005; Cools and D'Esposito 2011). In particular, a central role is played by the prefrontal cortex and its interactions with the rest of the brain, especially the specialized processing modules of the posterior cortex, including parietal (spatial attention) and inferotemporal (feature attention) areas; the declarative memory systems in the temporal lobes, including the rhinal cortex (recognition memory) and hippocampus (scene/episodic memory); and the language processing modules such as Wernicke's area, specialized in the comprehension of speech, and Broca's speech and production area. In addition, the prefrontal cortex interacts with subcortical structures such as the limbic structures involved in the processing of motivational and emotional cues as well as the orchestration of behavioral, autonomic, and endocrine responses, including the amygdala, hypothalamus, and brain stem centers; the basal ganglia, which are involved in the higher-order control of thought (see below) and action; and the neuromodulatory systems of the reticular core of the brain, including monoamine and cholinergic cell groups in the midbrain and hindbrain.

## Working Memory Models

As an example of how these interactions could support a balance between stability and flexibility, let us consider the case of working memory. In models of working memory, stability (i.e., stable goal-oriented performance) can be maintained by holding temporally stable representations of our goals. Goals for action can reside at different levels of task abstraction and unfold over different timescales. Importantly, however, a goal held in working memory can include the goal of meeting the requirement of specific tasks. As such, our ability to hold this goal in memory, available for use as a control signal, allows for stable task performance. Likewise, our ability to update working memory (i.e., to shift goals as context demands) is important for flexibility.

The control of working memory is often conceptualized as a gate that is distinct from the memory store itself (Hochreiter and Schmidhuber 1997). Closing an input gate against distracting information prevents its access to working memory, keeping the current contents available as control signals; this gating function promotes stability. In contrast, opening the gate enables the updating of working memory and allows new contextual information to modify behavior; this gating function promotes flexibility.

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

Though different mechanisms for working memory gating have been proposed (e.g., Wang et al. 2004; Zhu et al. 2018), one influential model has focused on frontostriatal interactions (Frank et al. 2001). This circuit is schematized in Figure 13.7 (O'Reilly 2006). The corticostriatal model of working memory gating proposes that the prefrontal cortex supports information maintenance, whereas the striatum-pallidal-thalamic pathway implements gating by regulating what information is allowed in and out of working memory.

Based on this model, spontaneous, unwanted events could be experienced as intrusions when the gate to working memory is breached and the intrusion supplants ongoing working memory processes. Once this occurs, intrusive events serve as signals to drive other cognitive processes and actions. Thus, the integrity of the working memory gating is paramount for mitigating against intrusive experiences. For example, by preventing an unwanted experience from updating to working memory or by inhibiting their influence on output control signals, one could stop the negative cycle of behaviors that can result from intrusive experiences. These gating mechanisms could be global (like the fast, inhibitory mechanisms supported by the hyperdirect pathway that can affect multiple processes simultaneously) or selective, supported by both the direct and indirect pathways, schematized in Figure 13.7 as the Go and No-Go pathways. Coordination among multiple corticostriatal loops can also be a mechanism for working memory operations in separate prefrontal areas to carry out complex, sequential, and hierarchically structured tasks (for a review, see Badre and Nee 2018).

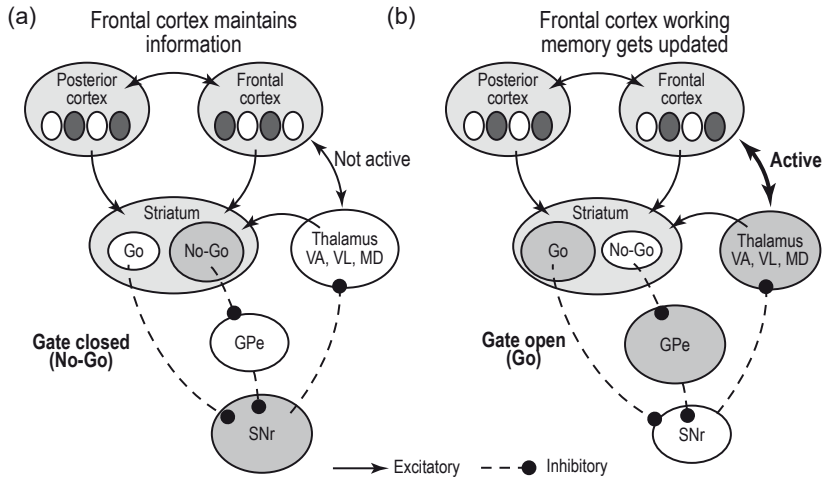



**Figure 13.7** Schematic depicting a mechanism of working memory gating through corticostriatal interactions. Inhibition (a) or disinhibition (b) of thalamocortical dynamics through the striatum can regulate gate closing and opening, respectively: VA (ventral area), VL (ventrolateral), MD (medial dorsal), GPe (globus pallidus external), SNr (substantia nigra pars reticulata). Reprinted with permission from O'Reilly (2006).

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 type of gating one selects can be thought of as a *gating policy*, in the same sense as defined above. As such, the match of the right gating policy to the particular dynamics of the situation is a key determinant of successful control (Bhandari and Badre 2018). For example, when confronted with an unwanted memory, one could deploy a global suppression to prevent it from entering working memory or instead attempt to selectively input another thought into working memory in its place. The consequences of these policies on memory or the ongoing impacts of the triggering event (both in this instance or in the future) might differ depending on the gating strategy that is selected. Thus, pathologies could arise as a result of any of the following:

- Items seeking to enter working memory are sufficiently salient or valued and will therefore breach the gating mechanism to update working memory (breach intrusion).
- Gating itself is weak and thus items access working memory, even if they are not adaptive or helpful to the individual (permissive intrusion).
- Mechanisms involved in maintaining stability (other than gating) are too strong and thus do not allow working memory to be updated once an intrusive experience has occurred.
- The wrong gating policy is selected given the nature or dynamics of the intrusion.

**Associative Learning Models**

Corticostriatal circuits are also central to *associative learning models* (Balleine and Dickinson 1998). Two control processes have been identified that are engaged in the control of goal-directed and habitual actions and which are mediated by distinct parallel circuits through the basal ganglia; in some circumstances, they compete with one another (Balleine et al. 2009). The goal-directed network is engaged rapidly with changes in the environment, incorporates the cortical working memory process described above, and utilizes this network to encode the action–outcome associations that mediate goal-directed action in a region of dorsomedial striatum. Generally speaking, this network relies on this prefrontal-dorsomedial-striatal (or caudate) pathway and feedback to the cortex via the substantia nigra pars reticulata and mediodorsal thalamus (Balleine and O'Doherty 2010) to encode and utilize novel solutions to problems presented by a changing environment. It also functions to inhibit older, more routine and outdated solutions, particularly the performance of habitual actions centered on the sensorimotor cortices and putamen or dorsolateral striatum (Graybiel 2008), when these have or are likely to have aversive consequences. If the goal-directed circuit is altered (e.g., through damage, disease, or drugs), inhibition can be reduced or mistimed, resulting in dysregulation of habits (even in the presence of aversive consequences),

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

and producing intrusive experiences. An increased reliance on habits may not only apply to the compulsive acts seen in OCD (Saxena et al. 1998; Robbins et al. 2019), but also the persistent motor habits (tics) associated with Tourette syndrome (Maia and Conceição 2017, 2018) as well as craving and compulsive drug use in addiction (Everitt and Robbins 2016; Furlong et al. 2017).

There are, however, other important features that are controlled by the goal-directed circuit, particularly by the dorsomedial striatal component of that circuit. As mentioned, considerable evidence suggests that the prefrontal working memory systems provide inputs to the striatum that mediate the plasticity necessary to encode goal-directed actions in the posterior segment of the dorsomedial striatum (reviewed in Balleine and O'Doherty 2010). However, to allow this large structure to encode more than one action–outcome association, plasticity associated with new action–outcome learning needs to be segregated from prior learning. It appears that this segregation is achieved via state-related information provided by inputs to the striatum from the parafascicular thalamus (Bradfield et al. 2013). This input onto the tonically active striatal cholinergic interneurons causes them to pause, allowing the principal neurons (the spiny projection neurons) relief from inhibition induced by tonic acetylcholine release. During this pause, cortical and midbrain dopaminergic inputs to the dorsomedial striatum can combine to induce plasticity in the spiny projection neurons. Accordingly, changes in action–outcome contingency provoke changes in the patterned input from the parafascicular thalamus, leading to plasticity changes in the targeted dorsomedial-striatal region.

Importantly, evidence suggests that the retrieval of specific action–outcome ensembles for performance is mediated by state-related information, based largely on outcome-related information (Bradfield et al. 2015) conveyed to the striatum, not by the parafascicular thalamus but via inputs from the orbitofrontal cortices (Gremel and Costa 2013; Bradfield et al. 2015; Stalnaker et al. 2016). Thus, accurate retrieval of specific action–outcome associations will be determined by the fidelity of this orbitofrontal cortical input: as a consequence, changes in orbitofrontal cortex activity (e.g., in OCD) could result in faulty retrieval, causing changes in flexibility (described above) and leading to the intrusion of unwanted information. Retrieval can become "frozen" if the orbitofrontal cortex gets "stuck" in a given state (see Appendix 13.1); alternatively, it could become highly, temporally disparate if activity in the orbitofrontal cortex fluctuates rapidly and unpredictably.

This type of state information features heavily in computational accounts, particularly model-based reinforcement learning accounts of goal-directed action. Such accounts provide information about state transitions for retrieval and could be seen as the computational implementation of these ideas (Wilson et al. 2014). See Appendix 13.1 for further computational modeling of recurrent intrusions in OCD focusing on neuromodulation within orbitofrontal cortex.

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

## Summary

Translating findings across different levels of analysis, including computational, psychological, neurobiological, and physiological, is challenging. However, to understand the nature of intrusive experiences and to develop effective treatments, such translation is essential. A first step in this translation is to use a common language. To this end, we have attempted to define all terms and concepts, especially where multiple, related, but somewhat distinct meanings exist. Salience is one such example of a term that is often used broadly to refer to the quality of being particularly noticeable, but in a Bayesian framework is specifically used to refer to the value afforded to uncertainty resolution. The models we discuss provide explanations for a range of intrusive experiences: from the obsessions and compulsions of OCD and drug and emotion-related intrusions in addiction and PTSD to intrusions of thoughts, bodily experiences, and mental images in anorexia nervosa. We focused on two major networks, frontostriatal and insula-cingulate, to illustrate how imbalances in these networks can lead to intrusive experiences. Whenever possible, overlap between different models and levels of analysis have been highlighted to provide a systems overview of how intrusive experiences across a range of distinct psychiatric, neurodevelopmental, and neurological disorders may emerge as a consequence of dysfunction at different levels of the nervous system.

## Appendix 13.1

### Active Inference

This appendix provides a technical description of belief updating under active inference. One useful aspect of treating "trains of thought" as "planning" under a generative model is that one can always express a generative model as a *graphical model* (Figure 13.A1). This is important because a graphical model can be used to understand the computational architecture of neuronal message passing in the brain. For every graphical model that specifies the states and outcomes in play and their conditional dependencies, there is an associated *factor graph* that provides, and must be supported by, unambiguous specifications of the architecture (e.g., neuronal connectivity) and message passing (e.g., neurophysiology); for details, see Figure 13.A2 and Friston et al. (2017).

In brief, the sorts of generative models commonly used to explain planning as inference are usually based on partially observed Markov decision process models. Crucially, in these generative models, discrete states of the world evolve over time in a way that *depends upon action*.

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 Free Energy, Salience, and Value**

For technical readers, expected free energy can be decomposed into an epistemic, information-seeking, uncertainty-reducing part (*intrinsic value*) and a pragmatic, goal-seeking, instrumental part (*extrinsic value*). Formally, the expected free energy for a particular policy $\pi$ at time $\tau$ in the future can be expressed as in terms of beliefs $Q(s_\tau, o_\tau | \pi)$ about future states $s_\tau$ and outcomes $o_\tau$:

$$G(\pi,\tau) = -\underbrace{E\left[\ln Q(s_\tau \mid o_\tau,\pi) - \ln Q(s_\tau \mid \pi)\right]}_{\textit{intrinsic value}} - \underbrace{E\left[\ln P(o_\tau)\right]}_{\textit{extrinsic value}}. \quad (13.A1)$$

Extrinsic (instrumental) value is simply the expected value of a policy defined in terms of outcomes that are preferred a priori. The more interesting part is the uncertainty-resolving or intrinsic (epistemic) value, variously referred to as relative entropy, mutual information, information gain, Bayesian surprise, intrinsic motivation, or value of information expected under a particular policy (Barlow 1961; Howard 1966; Optican and Richmond 1987; Linsker 1990; Itti and Baldi 2009).

Intrinsic (epistemic) value can be regarded as *salience*. Formally, this means that salience is the Kullback-Leibler (KL) divergence between posterior beliefs about hidden states with and without observations solicited by a particular act (or policy). The reason this divergence is associated with salience stems from the visual neurosciences, where the salience of a potential location for a saccadic fixation is known as *Bayesian surprise* (Itti and Baldi 2009; Sun et al. 2011; Barto et al. 2013). In robotics and machine learning, the information gain or Bayesian surprise is known as intrinsic motivation or value (Ryan and Deci 1985; Eccles and Wigfield 2002; Oudeyer and Kaplan 2007; Schmidhuber 2010; Barto et al. 2013). It is also referred to as *epistemic value* or *epistemic affordance* (Parr and Friston 2017). Epistemic affordance appeals to Gibsonian notions of affordance: it is the resolution of uncertainty afforded by a particular act: "What would I learn by looking over there?" On a psychological interpretation, intrinsic value can also be associated with *incentive salience* (Berridge and Robinson 1998; McClure et al. 2003). Exactly the same kind of mathematical arguments can be applied not just to beliefs about states in the world but also the parameters of the generative model. These parameters encode contingencies and laws governing the evolution of states or their mapping to observations. In this setting, salience becomes *novelty*; namely, the information gain afforded by knowing "what would happen if I did that?"

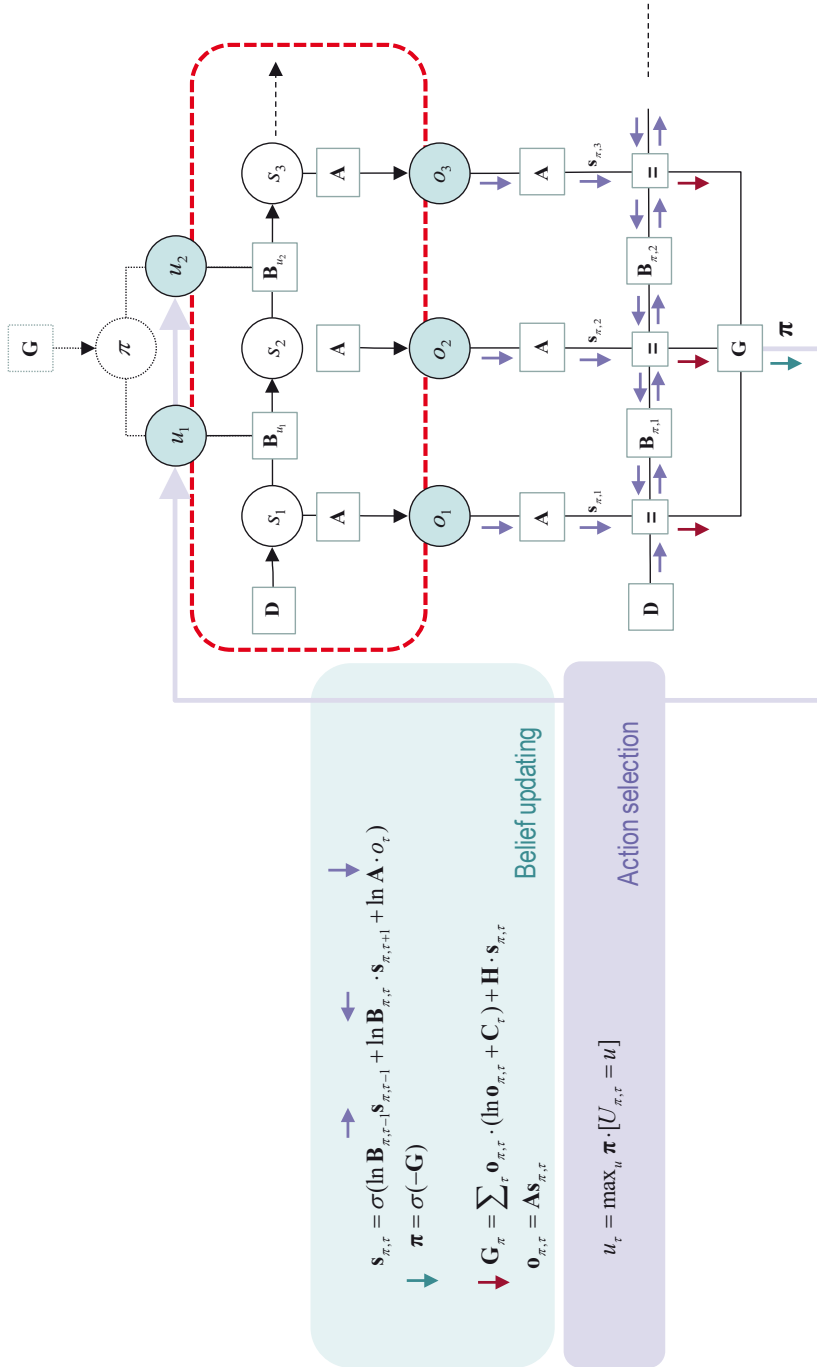
The factor graph in Figure 13.A2 is used to pass messages among the nodes (e.g., neuronal populations) to minimize free energy per se; in other words, to maximize the evidence for any given generative model of how outcomes were generated. This leads to biologically plausible message-passing schemes of the sort studied in terms of evidence accumulation and predictive coding

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

**Figure 13.A1** A generative model for discrete states and outcomes. Upper left panel: these equations specify a generative model. A generative model is the joint probability, *P*, of outcomes or consequences and their (latent or hidden) causes, see first equation. Usually, the model is expressed in terms of a likelihood (the probability of consequences given causes) and priors over causes. When a prior depends upon a random variable it is called an empirical prior. Here, the likelihood is specified by a matrix **A**, whose elements are the probability of an outcome under every combination of hidden states. *Cat* denotes a categorical probability distribution. The empirical priors pertain to probabilistic transitions (in the **B** matrix) among hidden states that can depend upon actions, which are determined by policies (sequences of actions encoded by $\pi$). The key aspect of this generative model is that policies are more probable a priori if they minimize the (time integral of) expected free energy **G**, which depends on prior preferences about outcomes or costs encoded in **C** and the uncertainty or ambiguity about outcomes under each state, encoded by **H**. Finally, the vector **D** specifies the initial state. This completes the specification of the model in terms of model parameters that constitute **A**, **B**, **C**, and **D**. Bayesian model inversion refers to the inverse mapping from consequences to causes (i.e., estimating the hidden states and other variables that cause outcomes). In approximate Bayesian inference, one specifies the form of an approximate posterior distribution *Q*. This particular form in this figure uses a mean field approximation, in which posterior beliefs are approximated by the product of marginal distributions over time points. Subscripts index time (or policy), italic variables represent hidden states, and bold variables indicate expectations about those states. Upper right panel: this Bayesian network or graphical model represents the conditional dependencies among hidden states and how they cause outcomes. Open circles are random variables (hidden states and policies), filled circles denote observable outcomes, and squares indicate fixed or known variables, such as the model parameters. Lower left panel: these equalities are the belief updates mediating approximate Bayesian inference and action selection. Lower right panel: this is an equivalent representation of the Bayesian network in terms of a Forney or normal style factor graph. Here the nodes (square boxes) correspond to factors and the edges are associated with unknown variables. Filled squares denote observable outcomes. The edges are labeled in terms of the sufficient statistics of their marginal posteriors (see approximate posterior). Factors have been labeled intuitively in terms of the parameters encoding the associated probability distributions (on the upper left). The circled numbers correspond to the messages that are passed from nodes to edges (the labels are placed on the edge that carries the message from each node). These correspond to the messages implicit in the belief updates (lower left). This figure is based on Friston et al. (2017).

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

**Figure 13.A2** The generative process and model. This figure reproduces the Bayesian network and Forney factor graph of Figure 13.A1; however, here the Bayesian network describes the process that generates data, as opposed to the generative model of data. This means that we can link the two graphs to show how the policy half-edge of Figure 13.A1 couples back to the generative process (by generating an action that determines state transitions). The selected action corresponds to the most probable action under posterior beliefs about action sequences or policies. Here, the message labels have been replaced with little arrows to emphasize the circular causality implicit in active inference: the real world (red box) generates a sequence of outcomes that induce message passing and belief propagation to inform (approximate) posterior beliefs about policies (that also depend upon prior preferences and epistemic value). These policies then determine action, which generates new outcomes as time progresses, thereby closing the action perception cycle. This figure is based on Friston et al. (2017).

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

(Srinivasan et al. 1982; Rao and Ballard 1999; Huk and Shadlen 2005; Beck et al. 2008; Bastos et al. 2012; Egner and Summerfield 2013; de Lafuente et al. 2015; Kira et al. 2015; Shipp 2016). In terms of the parameters of the generative model, associative plasticity is the corresponding belief update for neuronal connections (Friston et al. 2016).

## Precision and Parameters

Of particular interest here are the parameters that link states to outcomes and states at one point in time to states at the next point in time. In Figure 13.A1, these are simply matrices of probabilities encoding the likelihood mapping from states to outcomes **A** and the transition probabilities to one state to the next **B** (which depend upon a particular policy).

Neurobiologically, these matrices play the role of *connectivity* matrices, which play an important role in sensory data assimilation and subsequent planning based on beliefs about the consequences of any action. Furthermore, each column of these matrices has a *precision*. Precision, in this instance, reflects the fidelity or confidence about the outcome (or subsequent state) given the current state of the world. A very precise mapping means that we can be almost 100% confident that this will happen given that state, while a very imprecise mapping means that all outcomes (all subsequent states) are equally likely. For discrete space models, one can express the likelihood and priors in terms of inverse temperature or softmax parameters with the following form, where $\sigma(\cdot)$ is a softmax function or normalized exponential:

$$\begin{aligned} P\left(o_\tau \mid s_\tau\right) &= \sigma\left(\gamma_o \cdot \ln \mathbf{A}\right) \\ P\left(s_\tau \mid s_{\tau-1}, \pi\right) &= \sigma\left(\gamma_s \cdot \ln \mathbf{B}\right) \\ P\left(\pi\right) &= \sigma\left(-\gamma_\pi \cdot \mathbf{G}\right). \end{aligned} \tag{13.A2}$$

## Neural Organization of Interoceptive Processing

The control and representation of internal bodily physiology is instantiated throughout the neuraxis (for reviews, see Craig 2003; Critchley and Harrison 2013). While ganglionic and spinal reflexes support proximate physiological regulation, the brain orchestrates homeostatic control and allostatic responses across bodily organs, integrating control with behavioral demand. The brain receives interoceptive information about the internal state of the body via neural afferent and humoral interoceptive routes (for details, see Figure 13.A3). Somatosensory pathways also contribute to quasi-interoceptive sensation of bodily physiology (e.g., via heart beating against the chest wall, rib motion, pharyngeal airflow) and to referred pain (e.g., angina felt in shoulder). Chemosensory signaling (including $O_2/CO_2$, hormones, cytokines, blood pH, glucose, and hydration) occurs through central blood sampling at

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

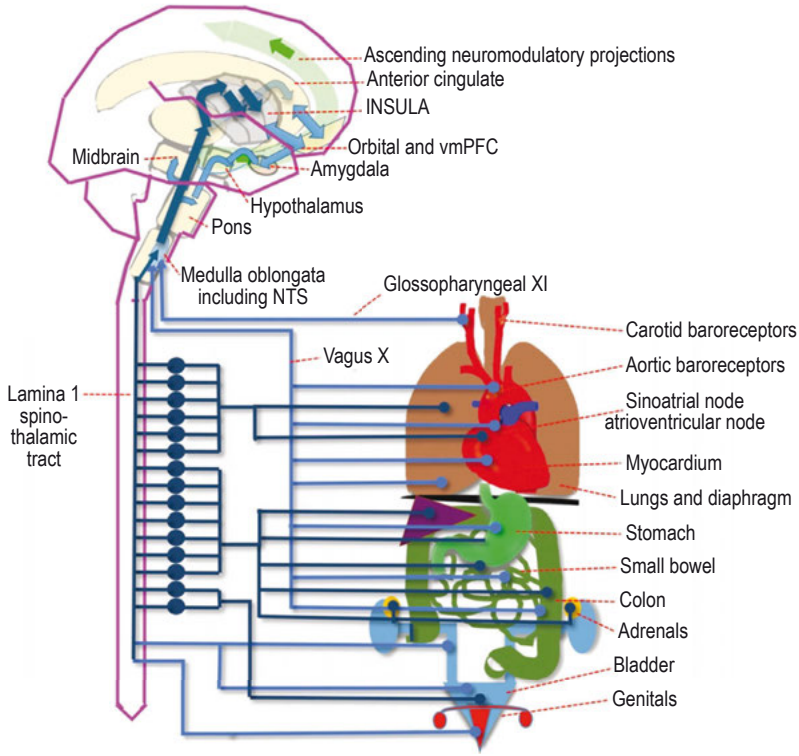



**Figure 13.A3** Schematic illustration of feed-forward neural interoceptive pathways. Peripheral afferents in cranial nerves X (vagus) and X1 (glossopharyngeal) and those following sympathetic nerves to spine (ascending laminar1) converge in the medullary nucleus of the solitary tract (NTS). Here, local connections support homeostatic reflexes (e.g., baroreflex) modulating autonomic outflow. Interoceptive information passes via a primary thalamocortical route to insula (shown in gray; viscerosensory cortex) where integrative processing builds a representation within anterior insular that is consciously accessible and can give rise to potentially intrusive interoceptive and affective feelings. Secondary interoceptive channels include (1) a subcortical route to hypothalamic and basal ganglia (including amygdala), modulating ascending widespread monoamine projections from midbrain (shown in light green) and (2) a thalamocortical route to visceromotor anterior cingulate cortex. Connections to orbitofrontal and ventromedial prefrontal cortices (vmPFC) offer another putative source of intrusive motivated feelings related to selection of action policies.

paraventricular organs and hypothalamus and may engender powerful motivational and arousal states (e.g., air hunger) with correspondingly intense feelings. The interoceptive representation within insular cortex shows a partial viscerotopy and connections follow a posterior-anterior and dorsal-ventral progression with increasing opportunity for cross-modal integration (Craig 2003, 2009; Evrard 2019). Anterior insula is most implicated in supporting conscious access to interoceptive sensations and associated emotional and motivational

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

feelings (Critchley et al. 2004), including the urge-to-tic in Tourette syndrome (Conceição et al. 2017) and drug cravings (Goldstein et al. 2009; Garavan 2010). Reciprocal connections between anterior insula and "visceromotor" rostral cingulate regions (both "allostatic" dorsal anterior cingulate and "homeostatic" subgenual cingulate) represent a putative functional architecture for higher-order predictive regulation of bodily states (Critchley et al. 2004; Critchley et al. 2005; Medford and Critchley 2010). As described above, anterior insula and dorsal anterior cingulate are key hubs within the so-called salience network (highlighting the motivational primacy of interoception, where salience is the epistemic value afforded by uncertainty resolution of Bayesian surprise; see above and Fedota and Stein, this volume).

## Example of a Computational Model to Explain Recurrent Intrusions in Obsessive-Compulsive Disorder

In addition to being intrusive, obsessions in OCD are both recurrent and "sticky" in the sense that they are difficult to shake from mind. From a dynamical systems perspective, these characteristics seem to suggest that obsessions correspond to attractors (Rolls et al. 2008; Maia and McClelland 2012; Rolls 2012; Maia and Cano-Colino 2015); that is, states toward which a system tends and from which it may have difficulty escaping.

OCD prominently involves disturbances in the orbitofrontal cortex and connected regions (Maia et al. 2008). Neurochemically, OCD may be associated with low serotonin and/or high glutamate. A biophysically detailed computational model of serotonin and glutamate modulation of the orbitofrontal cortex showed that both low serotonin and high glutamate tend to create excessively strong attractors in the orbitofrontal cortex (Figure 13.A4). The network tends to fall into these attractors and then has difficulty escaping from them. This is consistent with the perseverative responding to a previously rewarded visual stimulus displayed by marmoset monkeys following depletions of serotonin in the orbitofrontal cortex, either following reversal (Clarke et al. 2006) or extinction (Walker et al. 2009) of the association between the stimulus and reward.

In these simulations, neuronal activity was elicited by "manually" activating subsets of neurons. A more complete design would also have to incorporate the endogenous gating of information into (and out of) this local network, as was described above in the context of working memory. In addition, there are complex interactions between neuromodulatory levels and their effects across interacting brain structures. For example, the extent to which a monkey displays perseverative responding depends not only on low levels of serotonin in the orbitofrontal cortex but also on high levels of dopamine in the striatum (Groman et al. 2013). Moreover, alterations in these neuromodulators at the level of the orbitofrontal cortex can have profound opposing influences on the levels of the same or different neuromodulators in other structures, including the striatum (Roberts et al. 1994; Clarke et al. 2014) and the amygdala

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

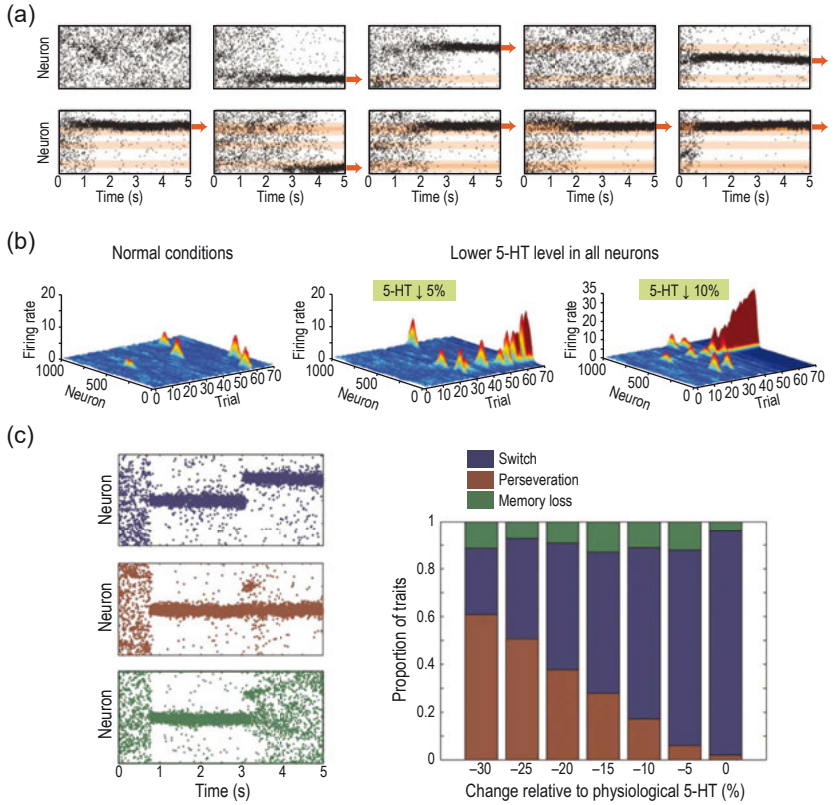



**Figure 13.A4** A computational model of the role of serotonin (5-HT) in the orbitofrontal cortex in OCD, adapted after Maia and Cano-Colino (2015). (a) Illustration of the process of entrenchment of patterns of neuronal activity, taken to correspond to obsessions or, in less pathological cases, "habits of thought." Each plot represents a population of neurons; the dots along each line represent the action potentials for one neuron. The network stochastically develops patterns of activity ("bumps"). Each bump elicits strengthening of the synapses between the neurons that were active in that bump through Hebbian learning, thereby developing attractors (orange bands). The more frequently a bump occurs, the more likely it is that it will reoccur (see last three plots). (b) Effects of reducing serotonin on the tendency to develop and fall into excessively strong attractors. Under normal circumstances, the network develops bumps at varying places over time (left panel). Under low levels of serotonin, however, the network tends to develop excessively strong attractors into which it repeatedly falls (middle panel). Moreover, there is a dose-response effect, such that reducing serotonin further causes even stronger attractors to develop (right panel). Increasing glutamate has the same effect as decreasing serotonin (not shown). (c) Low levels of serotonin cause the attractors to become excessively stable. Simulated activation of a set of neurons elicited a bump, followed by activation of a different set of neurons. Under normal conditions, the network's pattern of activity flexibly shifts to the state represented by the new bump (blue). Under low levels of serotonin, the network fails to shift to the new bump, resulting in perseverative activation of the prior bump (brown). (continued on next page)

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

**Figure 13.A4 (continued)** Importantly, low levels of serotonin increase such perseverative errors (brown) without affecting a different type of error in which the network simply loses the memory of what it was initially representing (green). The latter error, which is more reminiscent of disorders in which there is difficulty in keeping items in working memory (e.g., ADHD), is not affected by the serotonin manipulations.

(Roberts and colleagues, unpublished), which may exacerbate the possibility of intrusions occurring or becoming sticky. Understanding these effects and their implications for obsessions, if any, will require more complex models that incorporate the interactions between various regions and neuromodulators.

## Acknowledgments

Figures 13.A1 and 13.A2, from Friston et al. (2017), are used under the terms of the Creative Commons Attribution 4.0 International License (CC-BY).

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