10

Networks Relevant to Psychopathology and Intrusive Thought

John R. Fedota and Elliot A. Stein

Abstract

Intrusive thoughts are regular occurrences in healthy cognition. Across a variety of psychiatric conditions, however, such thoughts can become unconstructive and perseverative. Failures in computations to estimate the salience of the content of these thoughts are at least partly responsible for these clinically relevant disease symptoms. This chapter reviews neuroimaging results that show specific and related dysfunction in the calculation of salience at multiple neuroanatomically and functionally linked regions of interest, both cortically and subcortically. Transdiagnostic evidence for dysfunction in the striatum, thalamus, and prefrontal cortex is reviewed, as is a theoretical framework placing these regional findings in the context of large-scale brain networks. It is argued that changes in nodal function and network communication are signatures of a failure to properly shape predictions about the reliability and utility of external and internal stimuli, leading to maladaptive attentional capture and behavior, including intrusive thoughts.

Introduction

Spontaneous thought that is unrelated to current task demands is a normal part of healthy cognition. According to estimates from thought probe experiments, 30–50% of waking cognition is unrelated to specific tasks (Kane et al. 2007; Killingsworth and Gilbert 2010). However, even in normal cognition, this frequent mind wandering appears to bear an emotional cost. When prompted during mind wandering, participants reported being less happy than during task-focused cognition (Killingsworth and Gilbert 2010).

Spontaneous thoughts rarely occur only once. Instead, they are often recurrent. Across all types of repetitive thought, a common feature is an internal focus on "one's self and one's world" (Segerstrom et al. 2003). The content

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

of these internally focused thoughts can be either positive or negative, such as daydreaming versus worry (Watkins 2008; Christoff et al. 2016). An extensive taxonomy of repetitive thoughts exists (Watkins 2008), which describes both constructive and unconstructive consequences of intrusive and repetitive thoughts that can be linked to psychiatric conditions.

The ubiquity and diversity of repetitive thoughts during cognition beg the question: What characteristics mark the transition from healthy, spontaneous thought to clinically relevant, repetitive intrusive thought? It appears that the valence of the repetitive thought is an important factor in determining clinical relevance. In Watkins's (2008) taxonomy, negatively valenced thoughts with unconstructive consequences include depressive rumination, worry, and perseverative cognition. These consequences can clearly be linked to psychiatric conditions such as anxiety disorders, major depressive disorder (MDD), obsessive-compulsive disorder (OCD), posttraumatic stress disorder (PTSD), substance use disorders (SUDs), and schizophrenia (SZ). Negative repetitive thoughts (NRTs) become clinically relevant when their magnitude or frequency increases or when they become perseverative and difficult to control or eliminate (Kalivas and Kalivas 2016).

Cognitive Constructs Relevant to Negative Repetitive Thoughts

Beyond a taxonomy of NRTs, a description of the cognitive construct(s) underlying these clinical phenomena is necessary. Here, the goal is to operationalize the definition of NRTs as an intermediate step to identifying specific brain regions, circuits, and large-scale networks where the identified constructs are neurobiologically instantiated.

Cognition, intrusive or not, can be conceived as the making of predictions about the environment, testing those hypotheses via sensory processing, and subsequently updating and refining these predictions based on experience. Thus, optimal interaction with the world requires accurate beliefs about the environment and the ability to update those beliefs as new *reliable* evidence is encountered. This process can be computationally formalized as Bayesian predictive coding, a general theory of brain function that specifies how goal-directed behavior is motivated through the integration of prior beliefs with sensory information (Rao and Ballard 1999; Doya et al. 2007; Itti and Baldi 2009; Friston et al. 2012; Aitchison and Lengyel 2017). This framework integrates disparate cognitive processes including learning, reward, executive control, attention, and sensory processing.

Briefly, Bayesian predictive coding involves the integration of prior beliefs and available sensory evidence to refine posterior estimates of beliefs. Any mismatch between these two distributions signals the need to update beliefs considering the evidence encountered. The computational weight given to probabilistic estimates of either prior beliefs or sensory evidence is

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

governed by the relative precision of each (Mathys et al. 2011). That is, precisely estimated priors are less susceptible to modulation based on sensory evidence, while poorly estimated (imprecise) priors are readily adjusted by sensory evidence. As different sensory inputs can vary in their precision and information content, one cognitive imperative is to estimate which available sensory input will provide reliable and informative information to calculate better posterior estimates of the environment (Parr and Friston 2017, 2019). To this end, unambiguous sensory data should be amplified when present and sought when absent.

Cognitively, this predictive estimation of potentially reliable sensory sources has been described as the attribution of salience (Parr and Friston 2019). By salience, we mean a quality that is particularly noticeable or deemed important to the individual in a given context (Uddin 2014; Kahnt and Tobler 2017; Miyata 2019). As such, elements that provide unambiguous sensations should be ascribed higher salience, as they will provide reliable information for the adjustment of posterior estimates and iteratively improve subsequent decisions. The degree to which new information alters posterior estimates as compared to prior beliefs is termed *Bayesian surprise* (Itti and Baldi 2009; Barto et al. 2013).

Salience is closely related to value, but a key distinction is that value is a signed currency that varies monotonically from negative to positive whereas salience is an unsigned currency, where both negative and positive predicted outcomes can have equivalent salience (Kahnt and Tobler 2017). This is because both positively and negatively valent elements of the environment can improve posterior estimates; each can be informative in the refinement of posterior estimates.

When working properly, this iterative cycle of hypothesis (prior)–result (sensory evidence)–conclusion (posterior) allows for the flexibility and learning characteristics of human cognition. However, in the case of NRTs, improper predictions of the salience of elements in the environment will lead to suboptimal processing, including perseveration on elements with overly precise priors and/or a failure to guide attention to elements with weak priors.

For example, if previous experience creates an overly precise prior belief about the reliability of information to be gleaned from a stimulus (e.g., a drug cue or emotionally valent memory), its salience will be increased. Such an increase in anticipated salience will lead to the focus of greater attentional resources on the stimulus, potentially at the expense of alternatives in the environment. This dysregulated focus on one thought at the expense of others is a hallmark of intrusive thought, as described in the following sections. Indeed, the computational framework of Bayesian predictive coding has been shown to be useful in describing specific deficits across a variety of psychiatric conditions associated with NRTs: hallucinations in SZ (Sterzer et al. 2018), drug cravings in SUD (Gu and Filbey 2017), and perseverative focus in OCD (Levy 2018).

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

Below, we evaluate neuroimaging data across nodes and networks previously implicated in the cognitive processes related to the estimation of prior probabilities (salience) and attentional modulation within the framework of Bayesian predictive coding. Not surprisingly, the multiple points of failure in the information processing cascade requires discussion of a wide range of implicated brain regions. In addition, these individual brain regions can be supra-ordinately organized as nodes in large-scale networks, providing a systems-level perspective on dysfunctional communication associated with the estimation of salience. Each source of potential salience attribution dysfunction will be addressed in turn.

Ventral Striatum and Potentiated Response

A large body of literature has shown the striatum to be responsive not only to value and reward, but also to the salience of a given stimulus, independent of its positive or negative valence. In humans, these reward (magnitude * signed valence) and salience (magnitude * *absolute value* of valence) responses are separable: salience-evoked activation is seen in the ventral striatum (Zink et al. 2006; Jensen et al. 2007; Bartra et al. 2013), the insula, and the dorsal anterior cingulate cortex (dACC), while reward (positive valence) is also encoded in the striatum and across various brain regions, including the orbitofrontal cortex (OFC) (Litt et al. 2010; Bartra et al. 2013; Kahnt et al. 2014).

Dysfunction in striatal signaling related to the identification of salient environmental stimuli is observed across psychiatric conditions associated with NRTs. When compared to healthy individuals, those at risk for developing psychosis show heightened activation to task-irrelevant stimuli within the ventral striatum (Roiser et al. 2012; Schmidt et al. 2016). The observed increase in ventral striatal activation to irrelevant stimuli suggests an oversensitivity to uninformative cues in the environment. Within the Bayesian predictive coding framework described above, during normal cognition these irrelevant cues should be ascribed reduced salience, as they provide little to no sensory information with which to refine posterior estimates of the task environment.

Similar biases in ventral striatal activation are seen in other conditions such as OCD and bias in processing losses (Jung et al. 2011), MDD and anhedonia (Whitton et al. 2015), as well as SUD and cue reactivity (Volkow et al. 2010; Kühn and Gallinat 2011). Further, in a recent review of SUD, Moeller and Paulus (2018) observed that ventral striatal activation patterns are related to long-term abstinence outcomes: increased activation in response to drug cues is related to worse clinical outcomes (i.e., increased substance use and recidivism), whereas increased activation to monetary or nondrug reward cues are related to better clinical outcomes. In each case, ventral striatal activation is biased in its sensitivity to a variety of environmental cues. Thus, 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

dysregulated salience attribution does not appear to be a consistent, transdiagnostic insensitivity to reward or aversive stimuli. Instead, it appears there is an inability to distinguish properly between task-relevant and task-irrelevant (or intrusive) cues.

However, the question remains: How does such an increase in evoked activity precipitate a perseverative or intrusive thought? One well-articulated neurobiological mechanism, instantiated in the ventral striatum, for such a transition to salience misattribution and uncontrolled attentional capture is the glutamate-mediated transient synaptic potentiation model (reviewed by Kalivas and Kalivas 2016).

Briefly, this model suggests that upon presentation of a previously potentiated cue (e.g., a drug-related cue in SUD or an object of obsession in OCD), excessive glutamine release in the nucleus accumbens (NAc) core leads to transient synaptic potentiation that biases the attribution of salience to the potentiated cue being processed. In addition, this bias in the NAc core leaves the region less responsive to alternative cues that would normally be more fully encoded (Kalivas and Kalivas 2016). This dual mechanism likely increases the magnitude of the difference between potentiated and alternative cues and further instantiates salience of the potentiated cue.

The molecular mechanism for such a transient biasing of NAc core processing has been described in a rodent self-administration model: Drug-seeking behavior was related to transient potentiation of D1 receptors in the NAc core following presentation of drug cues. Moreover, following brief access to cocaine this potentiation rapidly extinguished, only to be reinstated following 45 minutes of forced abstinence (Spencer et al. 2017). These dynamics of potentiation are consistent with models of SUD that describe preoccupation with and craving for drugs, both of which can be viewed as NRTs (Koob and Volkow 2009).

Once a cue is ascribed a salience incommensurate with its task relevance, transient synaptic potentiation within the ventral striatum may sustain this bias by both increasing the coded salience of the potentiated (i.e., maladaptive) cue, while simultaneously decreasing the salience of any alternative, competing (i.e., goal-directed) cue. Computationally, this cascade is consistent with biased processing of specific subsets of previously encountered stimuli (e.g., drug cues in SUD), potentially leading to a reduced ability to modify posteriors based on experience. Maladaptive ventral striatal activation across a variety of psychiatric conditions is consistent with this interpretation.

Cortico-Striatal-Thalamo-Cortical Loops Convey Salience Signals throughout the Brain

Regardless of how the NAc core encoding of salience is biased in disease, neuroanatomical evidence clearly shows processing in the ventral striatum does not occur in isolation. Dense and reciprocal interconnections between

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

striatum, thalamus, and cortex in the form of cortico-striatal-thalamo-cortical (CSTC) loops allow for complex communication among brain areas when processing salient and motivationally relevant information. These connections also provide an anatomical mechanism to convey reciprocally looping ascending (striatum–cortex) and descending (cortex–striatum) influences on salience computation. The physical connections along CSTC loops have been described in great detail (Haber and Knutson 2009; Haber 2016), while tract-tracing results in nonhuman primates can be clearly related to patterns of resting-state functional connectivity in humans (Parkes et al. 2016; Choi et al. 2017).

Classically, three loops linking the dorsal striatum with presupplementary motor area, frontal eye fields, and dorsolateral prefrontal cortex (dlPFC), and two loops linking the medial and ventral striatum with the OFC and ACC, respectively, have been defined (Alexander et al. 1986). These connections form a gradient in cortical projections to the striatum, whereby ventral striatal inputs are associated with PFC areas processing emotion, caudate inputs with cognition, and putamen inputs with sensorimotor processing (Haber 2016). These interconnections suggest that the biased signals processed in the ventral striatum are ultimately conveyed throughout the brain, including specific regions discussed below. Many of these are primary nodes of large-scale networks across the cortex relevant to attentional control (Uddin 2014; Heilbronner and Hayden 2016) and psychiatric disease (Menon 2011; Sutherland et al. 2012; Kaiser et al. 2015).

However, it is important to note there is no clear anatomical or functional boundary between the ventral and dorsal striatum, and the cortical projections to and from the striatum create a continuum of connectivity (Haber and Knutson 2009; Choi et al. 2017; Marquand et al. 2017). Thus, assigning a one-to-one connectivity relationship between striatal and cortical regions is not possible. In fact, it has been estimated that the terminal fields of dACC, ventromedial PFC, and OFC cover almost 25% of the striatum, which is an overrepresentation as compared to the cortical volume of the brain (Haber et al. 2006). This overrepresentation is especially germane to the current discussion due to the role of OFC and dACC in reward and salience processing, respectively (Bartra et al. 2013). Instead of a well-defined gradient of CSTC loop connections, clear zones of integration as well as convergence are observed across the striatum (Haber and Behrens 2014; Choi et al. 2017; Marquand et al. 2017).

Choi et al. (2017) describe a homology between tract tracing in nonhuman primates and resting-state functional connectivity in humans, illustrating clear delineations in the pattern of physical and functional connection between the ventral and dorsal striatum. This pattern of connectivity agrees with previous tract-tracing evidence in nonhuman primates (Chikama et al. 1997). Specifically, more ventral striatum is strongly connected to the dorsal anterior insula (aIns) (Chikama et al. 1997) and dACC (Kunishio and Haber 1994; Parkes et al. 2016). In contrast, the rostral dorsal caudate is a hub 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

integration, with projections to and from a variety of brain regions associated with attentional control, including caudal inferior parietal lobule (IPL), ventrolateral, dorsolateral, and dorsomedial PFC, dACC, and OFC (Choi et al. 2017; Marquand et al. 2017).

The observed dichotomy in striatal projections to cortex is consistent with circuits differentially impacted by chronic exposure to drugs. Our group has identified a *ventral* striatal–dACC circuit whose resting-state functional connectivity is reduced in chronic cocaine users and a *dorsal* striatal–dlPFC circuit whose connectivity is increased in the same group as compared to healthy controls (Hu et al. 2015). Further, the balance between the up- and downregulation of these circuits was correlated with DSM-IV-TR compulsivity symptoms in cocaine users. Similarly, in first-episode SZ, reductions in ventral striatal–dACC resting-state functional connectivity are observed and have been correlated with reported symptom severity (Lin et al. 2017). These findings show that ventral striatal–dACC connectivity modulations are relevant across multiple conditions associated with NRTs (in this example, SUD and SZ). Further, the relationship between connectivity strength and clinically diagnostic criteria suggests a role for these circuits as potential neuroimaging biomarkers of disease severity.

The CSTC loops between cortex and striatum are interspersed with connections from striatal regions to various thalamic nuclei, which in turn are connected back to the cortex. The cortical and thalamic connections to the striatum are coordinated, meaning interconnected cortical and thalamic regions both project to the same striatal area. For the purposes of this discussion, the central-medial (CM) and medial parafascicular (PF) nuclei of the thalamus, which are connected to medial PFC areas including dACC (Behrens et al. 2003), are also connected to the ventral striatum (Van der Werf et al. 2002).

An illustrative example of impairment across nodes of the described CSTC loops in the processing of highly salient stimuli is seen in an imaging paradigm that employs erotic pictures (Metzger et al. 2010). When the salience of an anticipated erotic picture is processed by healthy participants, the nodes of this CSTC loop linking dACC and ventral striatum via the CM/PF thalamus become activated, as identified via high field fMRI. Here, CM/PF thalamus, dACC, and aIns showed increased activation during the anticipation of a salient, erotic picture (the ventral striatum was outside of the field of view in this study; Metzger 2010). Gola et al. (2016), provide further supportive evidence by showing enhanced anticipatory processing of erotic pictures within the ventral striatum of men seeking treatment for problematic porn use. Treatment seekers in this later study displayed enhanced striatal activation to the cues predicting erotic pictures but not to cues predicting monetary gains.

Taken together, Metzger et al. (2010) and Gola et al. (2016) show that anticipation, not consumption, of highly salient stimuli increases activation in each node of the CSTC loop (i.e., ventral striatum, CM/PF thalamus, and dACC). The observed activation within these nodes during *anticipatory* processing is

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

consistent with the conceptualization of salience as a predictive computational process for encoding anticipated relevance of a specific stimulus in the environment (Parr and Friston 2019), in this case an erotic image. These results provide an additional example of biased salience processing in a behavioral addiction (Potenza 2015; Kraus et al. 2016). This further broadens the scope of conditions associated with NRTs and dysfunction in core nodes of these CSTC loops.

The role of the thalamus in health and disease is an area of active inquiry. The traditional designation of the thalamus as a passive "relay station" for information processed elsewhere in the brain is being reconsidered. Recent evidence suggests thalamic influence on cortical connections, including the coordination of activation and direct control of synchronicity between cortical regions via gain control (Saalmann 2014), and active filtering of information (for a review, see Halassa and Kastner 2017). Consistent with this more active processing role, the mediodorsal thalamus is suggested to amplify signals in the PFC (Parnaudeau et al. 2018) and to extend representations in the PFC over longer time durations than those associated with cognitive processes, like working memory (Pergola et al. 2018). Advances in thalamic parcellation (Kumar et al. 2017) and quantification of its resting-state functional connectivity with the entire brain (e.g., via 7T fMRI) will only improve the resolution of these findings and further articulate the role of thalamic nuclei in salience attribution.

Prefrontal Cortex: Regions of Interest Implicated in Psychiatric Disease

With NAc core dysfunction now described along with neuroanatomical links between this region and thalamic and prefrontal regions, including the dACC and aIns, our focus shifts from striatal to cortical areas implicated in psychiatric disease associated with NRTs. A recent meta-analysis of structural ($n > 15{,}000$) (Goodkind et al. 2015) and functional ($n > 5{,}000$) (McTeague et al. 2017) data from psychiatric patients across a variety of disorders associated with intrusive thoughts (SZ, MDD, SUD, OCD) identified specific yet overlapping areas of dysfunction. Specifically, the pattern of gray matter loss in patients was circumscribed to dACC and bilateral aIns (Goodkind et al. 2015). In agreement with these structural findings, hypoactivation in cognitive control task-evoked activity was also observed in the dACC and right aIns as well as in the left dlPFC and right IPL (McTeague et al. 2017). The tasks employed in the meta-analysis did not probe intrusive thoughts per se, but rather top-down cognitive control more generally. That said, the pattern of hypoactivation across disease conditions showed reductions in regions associated with both attentional control (i.e, left dlPFC, right IPL) and the calculation of salience (i.e., dACC, aIns).

The aIns and dACC are coactivated across a wide variety of cognitive control and attention-related tasks. In fact, activation in these nodes are among the most observed results in the fMRI cognition literature (Behrens et al. 2013).

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

Given their ubiquity in the extant literature, aIns and dACC have been theorized to play a central role in broad cognitive constructs, as in the detection of relevant information from both the external and interoceptive worlds and the coordination of appropriate attentional capture and behavioral response to these salient signals (Uddin 2014; Heilbronner and Hayden 2016; Nour et al. 2018). The degree of task activation in these regions increases with the demand for attentional control or with an increase in ambiguity of stimuli during perceptual decision making (Lamichhane et al. 2016).

The insula is associated with the integration of interoceptive information into calculations of salience (Critchley et al. 2004; Craig 2009) via ascending pathways communicating visceral and allostatic information (Critchley and Harrison 2013; Kleckner et al. 2017). Within the insula (and the dACC, as discussed below), these signals are integrated to create a single subjective image of "our world" (Kurth et al. 2010); interoceptive representations in the insula have been theorized to provide the basis for a perception of self via the integration of interoceptive signals related to physical state (Seth 2013; Namkung et al. 2017). This self-focused processing localized to the insula is strikingly consistent with the description of the content of repetitive thought as being "related to one's self and one's world" (Segerstrom et al. 2003).

Returning to a transdiagnostic theme and the observation of insular hypoactivity (McTeague et al. 2017), impairments in interoception are implicated across psychiatric conditions associated with NRTs (Khalsa et al. 2018). Insular activation in response to cued recall of a previously interoceptive challenge was diminished in subjects with MDD as compared to healthy controls (DeVille et al. 2018). Similar impairment in interoceptive processing focused on insular hypoactivation (Naqvi and Bechara 2010) has also been associated with SUD (Goldstein et al. 2009; Sutherland et al. 2013; Paulus and Stewart 2014).

In addition to a central role in the integration of interoceptive signals, the insula is a primary adjudicator between external (exogenous) and internal (endogenous) focus as a function of salience attribution (Sridharan et al. 2008; Menon and Uddin 2010; Uddin 2014). Characterization of the causal interactions between insula and dACC, however, shows additional divergence in their patterns of activation. Specifically, aIns has been shown to amplify the detection of salience within the dACC (Chen et al. 2014; Cai et al. 2015). This timescale is consistent with EEG spectral analysis showing insular activation precedes that of dACC (Chand and Dhamala 2016).

In contrast to the association with interoceptive processing in the insula, the dACC is associated with cognitive monitoring and control processes along with economic decision making (Botvinick 2007; Kolling et al. 2016; Shenhav et al. 2016; Alexander and Brown 2017). Recent integrative accounts suggest that the primary function of the dACC is to process multiple facets of information about the context in which a decision is being made to enable the appropriate goal-directed strategy (Heilbronner and Hayden 2016; Li et al. 2018b). That is, dACC codes task-state information (originating either endogenously

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 exogenously) that is relevant to the current demands on the individual via adaptive coding to discount or ignore irrelevant details (Heilbronner and Hayden 2016). An important difference between these context-relevant calculations and those occurring in the OFC, which are more closely linked with reward, is that while the OFC encodes the value of the current choice, the dACC integrates across multiple dimensions and across a longer timescale (Kennerley et al. 2011).

The related calculations of interoceptive salience and context-relevant stimuli within the aIns and dACC can be more formally combined via recent studies of belief updating in healthy populations. The belief updating calculation incorporates the integration of new, relevant information with existing expectations, termed Bayesian surprise (Barto et al. 2013). Divergence between prior belief and updated posterior beliefs characterizes the level of this "surprise." Recent work in healthy individuals shows that surprising (salient) information *relevant to updating beliefs* is encoded in the ventral striatum, aIns, and dACC whereas merely surprising but uninformative information is not (Nour et al., 2018). This is an important distinction, as these regions appear to differentiate the information content of different sensory streams. This function is in line with the construct of salience, as defined by Parr and Friston (2019): the anticipated reduction in uncertainty is associated with a specific element in the environment.

Thus, the increase in activation across the striatum, aIns, and dACC can be putatively related to the accurate identification of relevant information (i.e., the salience definition described above) used to refine beliefs and guide goal-directed behavior. Thus, along with the ventral striatum, the aIns and dACC play a combined role in distinguishing novel (unexpected, uninformative) and surprising (unexpected, informative) sources of information. Informative information is used to alter response strategies (e.g., attentional modulations or motor output), whereas surprising but irrelevant information is not.

Applying this idea more directly to psychiatric conditions that are linked to NRTs, a failure to accurately encode Bayesian surprise is consistent with the hypoactivations in aIns and dACC described by McTeague et al. (2017). Hypoactivation in these regions points to an inability to update beliefs efficiently, which in turn may perpetuate a positive feedback loop whereby poor estimates of salience guide attention to maladaptive elements in the environments, leading to perseverative focus on uninformative stimuli, which may precipitate NRTs.

The Salience Network in Intrusive Thought

Within the described CSTC loop, dysfunction and/or maladaptive bias toward potentiated stimuli is observed at multiple levels of salience processing. Additionally, many of the implicated regions of interest are regularly activated

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 concert depending on task demands. For example, Nour et al. (2018) show ventral striatal, dACC, and aIns activation in the processing of informative, novel information, with the dACC and aIns regularly coactivated (Behrens et al. 2013). These commonalities suggest a benefit to examining the brain at a higher level of organization, moving from regions of interest to dyadic circuits to large-scale network organization.

To this end, dACC and aIns are the primary nodes of the salience network (SN). Originally described by Seeley et al. (2007), and subsequently replicated using a variety of methodologies (Dosenbach et al. 2007; Smith et al. 2009; Power et al. 2011), the SN is a centralized processor that ascribes salience to stimuli (Uddin 2014) while coordinating attentional resources between an internal and external focus in response to task demands (Sridharan et al. 2008; Menon and Uddin 2010). Such a function is clearly consistent with the roles ascribed individually to aIns and dACC detailed above.

While the dACC and aIns form the primary nodes in the SN, additional large-scale networks are broadly relevant to cognition (Smith et al. 2009; Ji et al. 2019). The interaction between the SN and these networks leads to a more global view of the dysfunction associated with intrusive thought; such SN interactions can be conceptualized within a tripartite network model. Briefly, this model describes a default mode network (DMN), which includes the rostral and posterior cingulate cortices, parahippocampal gyrus, and bilateral inferior parietal cortex (Raichle et al. 2001; Buckner and DiNicola 2019), and an anticorrelated executive control network (ECN), which includes the bilateral dlPFC and parietal cortices (Honey et al. 2007). During interoceptive processing, the DMN is relatively more active and the ECN is relatively deactivated (Fox et al. 2005; Keller et al. 2013). During exteroceptive processing the reverse is true. Toggling between these two networks is thought to be mediated by the SN (Menon and Uddin 2010).

Indeed, regardless of the cognitive function ascribed to the individual regions of the SN, transdiagnostic findings implicating nodes of the SN, both structurally (Goodkind et al. 2015) and functionally (McTeague et al. 2017), have led to conceptualizations of SN dysfunction as a core transdiagnostic symptom of psychiatric dysfunction (Menon 2011). Transdiagnostic differences in activation within nodes of the CSTC loops connecting striatum to SN nodes have been described in detail by Peters et al. (2016).

Tripartite Network Model and Aberrant Cognition

An influential theoretical model by Menon (2011) describes a transdiagnostic hypothesis of aberrant salience calculation in psychopathology that centers on dysfunctional network connections between SN, DMN, and ECN. The model suggests that the normal adjudication between internal (DMN) and external (ECN) focus mediated by the SN (Sridharan et al. 2008) is commonly disrupted

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

across psychiatric diseases. While entirely consistent with biased calculations within the nodes of the SN (aIns and dACC) and ventral striatum described above, this network-level model provides a systems neuroscience description of brain dysfunction and explicitly relates changes in salience attribution to emotional and attentional processing distributed throughout the brain.

The dynamic interactions between large-scale networks are determined by a variety of factors. For example, the integrity of connections within a given network (Honey and Sporns 2008; Boes et al. 2015) is important to determine the functional capabilities of that network and the communication fidelity both within and between networks. In addition, high-fidelity messages can fail to be communicated successfully if hubs of interconnection between networks are dysfunctional in disease (Cole et al. 2013; Gratton et al. 2018). These communications also occur dynamically, and the engagement or disengagement of functional connectivity between networks, a process mediated by the SN, are potential points of failure at a systems level.

Menon's model of dysconnectivity has proven prescient (Menon 2011). Across a variety of diseases associated with NRTs, network-level dysconnectivity between SN–DMN, SN–ECN, and DMN–ECN have been observed. Using the tripartite network framework, a meta-analysis ($n > 16{,}000$ total patients, including $n > 8{,}000$ patients) across a variety of psychiatric diseases (including MDD, OCD, PTSD, SZ) associated with NRTs identified transdiagnostic patterns of seed-based functional dysconnectivity both within nodes of the SN, DMN, and ECN as well as between the three networks (Sha et al. 2019). Hypoconnectivity is observed at rest across diseases *within* nodes of the DMN and SN and *between* nodes of SN and both DMN and ECN. In contrast, hyperconnectivity is observed within the ECN and distinct nodes within DMN, between the ECN and DMN, and between DMN and subunits of the SN.

These results are broadly consistent with Menon's tripartite model: reduced connectivity between SN and DMN or ECN suggests a reduced complement of information to integrate into accurate and/or flexible estimates of prior and posterior beliefs about the environment. However, hyperconnectivity between DMN and nodes of the SN are also observed, suggesting the balance between connections, as opposed to a unitary increase or decrease between networks, may be a distinguishing feature (e.g., Hu et al. 2015).

A second recent result further broadens the scope of analysis: interrogating brain-wide connectivity patterns via a connectome-wide association study approach (Shehzad et al. 2014) to identify brain regions whose whole-brain connectivity pattern is associated with the p factor, a hypothesized common factor underlying psychopathology across disease conditions (Caspi et al. 2013). Using a data-driven approach and a large data set ($n > 600$), four regions in the occipital cortex were identified where whole-brain multivariate connectivity patterns were correlated with p factor scores (Elliott et al. 2018).

These results are not directly interpretable within the tripartite network hypothesis, as the regions identified fall outside of traditional SN, ECN, DMN

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

boundaries, when the identified occipital regions are used as seeds in a resting-state functional connectivity analysis, similar to those analyzed by Sha et al. (2019). Nonetheless, hyperconnectivity with both the ECN and DMN was positively correlated with p factor score. Importantly, while DMN and ECN within the tripartite network model are usually considered oppositional (Fox et al. 2005; Sridharan et al. 2008; Menon and Uddin 2010), the connectivity of both networks was similarly enhanced with an independent, and psychiatrically relevant, region of the occipital cortex. Further, the degree of this shared increase in functional connectivity correlated positively with p factor score.

These results point to network interactions beyond the SN (Peters et al. 2016) that may indirectly influence the salience calculations instantiated within dACC and aIns. In both of these recent cases, the network structure separating ECN and DMN appears to be reduced, either through hyperconnectivity between these two normally oppositional networks (Sha et al. 2019), or via increased coherence with a mediating node outside of the tripartite network (Elliott et al. 2018). The coherence between ECN and DMN may further bias the information integrated in SN, though an empirical demonstration of this remains outstanding.

Conclusion

In psychiatric conditions that include NRTs as a symptom, a common set of biases in information processing and dysconnectivity between specific nodes as well as large-scale networks is observed. In each case, these dysfunctions appear to reflect a failure to tune or modulate the response or connection properly, as opposed to a broad deficit in either task-evoked activation or connectivity. Recent advances in gathering large data sets across a diversity of psychiatric conditions have aided in revealing these dysfunctions.

It is important to note that *healthy* cognition includes the regular experience of intrusive thoughts. It is an increase in the perseverative focus on these thoughts that leads to clinically relevant dysfunction. Thus, it is not the presence of a response within or a connection between brain regions that is indicative of a disease as much as it is the inability to discriminate properly among alternatives or determine the most relevant information to guide decision making. We suggest that processing biases at the level of the striatum, thalamus, insula, and dACC indicate computational dysfunctions during the Bayesian predictive coding of salience.

As a representative example, we extend the model by Kalivas and Kalivas (2016) to incorporate a more explicit role for thalamus, dACC, and aIns along with large-scale networks within the tripartite model of brain function. Biased processing of potentiated stimuli, at the expense of alternative stimuli, within the ventral striatum is strongly linked to glutamate-mediated transient synaptic potentiation (Spencer et al. 2017). These signals are conveyed along 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

well-described CSTC loops (Choi et al. 2017) to the thalamus, where recent evidence suggests their representations in PFC may be amplified or sustained (Parnaudeau et al. 2018; Pergola et al. 2018), potentially increasing bias.

Hypoactivation of nodes within the SN (Peters et al. 2016; McTeague et al. 2017) leads to suboptimal integration of interoceptive signals (Kleckner et al. 2017), which may be biased due to the allostatic load of psychiatric disease more generally (McEwen and Gianaros 2011). In addition, the integration of endogenous and exogenous information to identify and fully process contextually relevant information (Heilbronner and Hayden 2016) is likely biased by the observed hypoactivity within nodes of the SN. These processing biases within SN lead to an inability to identify relevant information in a given context (Parr and Friston 2019) or to update beliefs to modify attentional strategies or behavior more generally (Nour et al. 2018). Finally, the integrative calculations of the SN may be further impacted by alterations in large-scale network structure (Sha et al. 2019), which further bias salience calculations by reducing the fidelity of information communicated with the rest of the brain.

In summary, bias at each stage of salience attribution leads to an overrepresentation of potentiated stimuli as well as to an insensitivity to counterfactual evidence, which normally signals the need to alter behavior. A better understanding of the calculations instantiated within each of these regions, and more importantly, a more holistic, systems-based picture of their interactions, is likely to identify novel therapeutic interventions that will allow us to mitigate the unconstructive consequences of NRTs and to treat the underlying dysfunction.

Open Questions

To guide future enquiry, we conclude by highlighting three problem areas that await clarification through future research. First, are NRTs the cause of the psychiatric diseases described or only a symptom? While the current conceptualization of failures in Bayesian predictive coding computations is consistent with the neuroimaging evidence of dysfunction in these conditions, few direct links have been described between these regions and the subjective experience of NRTs in patients (for a notable exception in SZ, see Sterzer et al. 2018). This computational framework, however, provides testable hypotheses to determine how the estimation and updating of beliefs about the environment may be causally linked to the experience of NRTs across conditions.

Second, what are the limits of the neurobiological framework centered on CSTC loops? Any discussion of salience necessarily implicates the entire brain. Which key nodes have not yet been accounted for in the current conceptualization (hippocampus, amygdala, dlPFC)? Especially in the estimation of beliefs, the central role of memory processes is currently underspecified.

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

Finally, what potential treatments do these circuits suggest? Given the multiple levels of systems that are impacted—from D1 receptors in ventral striatum (Roberts-Wolfe et al. 2018) to large-scale brain networks (Sha et al. 2019)—are multipronged treatments, such as simultaneous pharmacotherapy (Kalivas and Kalivas 2016) and transcranial magnetic stimulation (Peters et al. 2016) more likely to succeed?

Acknowledgments

This work was supported by the National Institute on Drug Abuse, Intramural Research Program and Center for Tobacco Products (U.S. Food and Drug Administration) Grant No. NDA13001-001-00000 (to EAS).

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