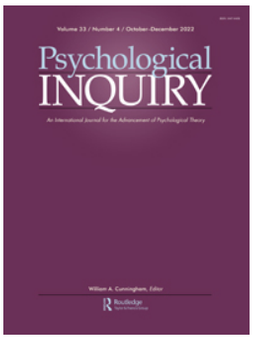

Routledge
Taylor & Francis Group

Psychological Inquiry

An International Journal for the Advancement of Psychological Theory

ISSN: 1047-840X (Print) 1532-7965 (Online) Journal homepage: www.tandfonline.com/journals/hpli20

Dynamics of Internal Attention and Internally-Directed Cognition: The Attention-to-Thoughts (A2T) Model

Iftach Amir & Amit Bernstein

To cite this article: Iftach Amir & Amit Bernstein (2022) Dynamics of Internal Attention and Internally-Directed Cognition: The Attention-to-Thoughts (A2T) Model, Psychological Inquiry, 33:4, 239-260, DOI: 10.1080/1047840X.2022.2141000

To link to this article: https://doi.org/10.1080/1047840X.2022.2141000

PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 4, 239–260
https://doi.org/10.1080/1047840X.2022.2141000

Routledge
Taylor & Francis Group

TARGET ARTICLE

Dynamics of Internal Attention and Internally-Directed Cognition: The Attention-to-Thoughts (A2T) Model

Iftach Amir and Amit Bernstein

University of Haifa, Haifa, Israel

ABSTRACT

The propensity to focus attention inwards is fundamental to human mental life and internally-directed cognition (IDC) [e.g., mindwandering, (mal)adaptive self-reflection]. Yet, understanding of the mechanisms through which internal attention shapes IDC is limited. We argue that understanding the systemic *complexity* and *dynamics of how internal attention interacts with other cognitive processes* can significantly facilitate our capacity to predict and model (mal)adaptive IDC. We, therefore, introduce the Attention-to-Thoughts model—a dynamic systems theory and computational model of internal attention in IDC. Through the model we aim to, first, conceptually and computationally define *momentary states* of this dynamic system; second, simulate and predict differential temporal *trajectories* of this dynamic system through which IDC emerges. Through experimental simulations, we explore how Attention-to-Thoughts may be used to better understand how internal attention selection is expressed from moment-to-moment; how internal attention unfolds by documenting how, as a function of contextual demands for focused attention, internal attentional selection iteratively transacts with working-memory and emotion; and, in turn, how maladaptive IDC (e.g., repetitive negative thinking, cognitive dyscontrol) emerges from temporal trajectories of the dynamic system of internal attention. Finally, we highlight key conceptual, computational, and methodological directions for the study of internal attention, IDC, and related phenomena (e.g., mindfulness).

KEYWORDS

Internal attention; internally directed cognition; mindwandering; dynamic systems; rumination

Introduction

Our propensity to direct attention inwards—onto our thoughts, memories, and feelings—is quintessential to human mental life and conscious experience. We spend as much as ~30–50% with our attention turned inwards, "looking in" (Killingsworth & Gilbert, 2010; Klinger, 1978), often also termed "internally-directed cognition" or "self-generated cognition," or simply *thinking* (Axelrod, Rees, & Bar, 2017). Internally-directed cognition (IDC) is an umbrella term for a number of inter-related mental phenomena, such as reflective attention, self-focused or self-directed attention as well as mindwandering, internal processing or mentation, imagery, and mental time travel (Axelrod et al., 2017; Chun & Johnson, 2011; Ingram, 1990). IDC is therefore a fundamental characteristic of human mental life and is thought to be functionally important (Klinger, 2013). Critically, internal attention determines what internal information reaches awareness, including the selection of thoughts, memories, images, and related internal experiences (Posner, 1994). Accordingly, whereas some of our looking in enable adaptive thinking (e.g., goal-setting, problem-solving, mental time travel), it can alternatively potentiate maladaptive thinking (e.g., spirals of negative repetitive thinking). Yet, our understanding of *how*, *when* and *for whom* (mal)adaptive IDC emerges from internal attentional selection, is limited and fragmented. Accordingly, we introduce the *Attention-to-Thoughts* (A2T) model—a dynamic systems theory and computational model of internal attention in IDC. Specifically, A2T characterizes how internal attention is expressed and iteratively transacts with lower-level working memory and emotion, from moment-to-moment in time, as a function of contextual demands on sustained focused attention, and thereby how (mal)adaptive thinking or internally-directed cognitive processes emerge and unfold.

External and Internal Attention

In an ongoing process of selection and modulation, our minds are challenged to process events in our (external) environment as well as the (internal) events within our mind including our thoughts, memories, and mental images (Klinger, 1978; Uddin, 2015). Due to the brain's limited processing capacity, not all concurrent external or internal events may be processed—certain information must be *selected* over competing information for preferential process-

CONTACT Amit Bernstein abernstein@psy.haifa.ac.il Observing Minds Lab, School of Psychological Sciences, University of Haifa, Mount Carmel, Haifa, 31905, Israel.

Table 1. Definitions of key concepts.

Term	Definition
Attentional selection	The preferential allocation of limited processing resources to certain sources of information over concurrently competing sources
Attentional modulation	The extent of facilitated processing of selected source of information
Internal attention	"Selection and modulation of internally generated information" (p. 77; Chun et al., 2011)
External attention	"Selection and modulation of sensory information" (p. 77, Chun et al., 2011)
Working memory	The limited capacity to store information in the mind in a highly accessible and modular state such that information can be quickly retrieved and manipulated

ing (*modulation*) (Chun, Golomb, & Turk-Browne, 2011; Desimone & Duncan, 1995; see Table 1).

External attention is the attentional processing of perceptual-sensory information incoming from various sources external to the mind/brain, such as the peripheral nervous system, originating from outside and/or with-in the body (e.g., visual information incoming via the eyes, proprioceptive sensations originating from the muscles). *Internal attention* is the attentional processing of information stored in the mind, whether recalled from long-term or active in working memory (Chun et al., 2011; Dixon, Fox, & Christoff, 2014; Gazzaley & Nobre, 2012). That is, internal attention biases processing in favor of certain internally generated- or stored- mental representations over other competing internal and external objects or stimuli (Myers, Stokes, & Nobre, 2017). Emerging work indicates that the goal-directed/stimulus-driven systems which govern *external-perceptual* processing may also operate over processing of *internal* events (e.g., memories, thoughts; (Chun et al., 2011; Chun & Johnson, 2011; Dixon et al., 2014; Uddin, 2015; Ziegler, Janowich, & Gazzaley, 2018).

Executive control processes, such as working memory and response selection, are by definition *internal and goal-directed* processes (Chun et al., 2011). Other forms of cognitive processes and states may be characterized as *internal and stimulus-driven* processes. For example, unwanted memories or involuntary remembering, or any interesting memory that "enters consciousness and takes over attentional resources" have been conceptualized as a result of automatic reflexive bottom-up forms of internal attention (Cabeza, Ciaramelli, Olson, & Moscovitch, 2008; van Schie & Anderson, 2017). This interplay between internal goal-directed and stimulus-driven systems is analogous to recent theory seeking to characterize dynamics of spontaneous thought processes including how thought processes, such as mindwandering, rumination and goal-directed thoughts arise and change over time (Christoff, Irving, Fox, Spreng, & Andrews-Hanna, 2016). Such thought processes are driven by dynamics between varying levels of deliberate (i.e., cognitive control/goal-directed system) and automatic (i.e., affective and sensory salience/stimulus-driven system) as well as contextual and mnemonic (e.g., working-memory capacity) constraints on thinking (Christoff et al., 2016).

Likewise, *selection history* may be especially meaningful for understanding internal attention and its role in IDC. In the external attention literature, a "selection history" system has been proposed to account for attentional effects, such as reward history, priming, or (previously learned) statistical regularities of how relevant and irrelevant information is distributed in the environment (Awh, Belopolsky, & Theeuwes, 2012; Theeuwes, 2019) Similarly, internal attention may also be biased to representations that had been previously selected (Theeuwes, 2019). For example, imagine a person who is preoccupied with thoughts regarding an on-going personal crisis or an unresolved problem. In attentional terms, this means that crisis-related thoughts (representations) are selected into WM and awareness, and the selection history of such representation begins to form or "accumulate" (Theeuwes, 2019). Internal selection history may bias continued and ongoing internal selection to related thoughts and memories of the crisis (Ehring & Watkins, 2008) thereby maintaining IDC (e.g., negative repetitive thinking) (Everaert, Bernstein, Joormann, & Koster, 2020).

Working Memory and Internal Attention

Although there has been limited study of *internal* attention per se (Chun et al., 2011; Kiyonaga & Egner, 2013; Nobre et al., 2004), there is extensive cognitive psychology and neuroscience focused on *working memory*. Working memory (WM) refers to the mental capacity to store and manipulate information in the mind (Baddeley, 2012; Myers et al., 2017). WM is inherently *attentional*—in that it is limited in capacity and so entails preferential processing of selected information. Critically, WM is also inherently *internal*—in that the information processed is independent of external sensory stimulation. Because working memory and internal attention are closely related processes (Kiyonaga & Egner, 2013; Myers et al., 2017; Oberauer, 2009), we use the term *internal attention* to refer to the specific processes of preferential selection and modulation of internal objects. In the present paper, when we use the term *working memory*, we refer mostly to its function in short-term storage and guiding behavior (Myers et al., 2017; Oberauer, 2009). We do so for simplicity and in line with models of WM that emerged from the short-term memory literature and theory distinguishing attentional from other processes (e.g., memory encoding, retrieval, or capacity) subserving WM (Baddeley, 2012; Oberauer, 2009).

Internal Attention, Working Memory, and Internally-Directed Cognition

Internal attention and WM serve critical functions in IDC. Attentional selection drives the gating or filtering of information that reaches conscious experience (Posner, 1994).

As such, attention functions to determine whether the focus of awareness and cognition is directed externally or inwardly (Verschooren, Schindler, De Raedt, & Pourtois, 2019) as well as which of the vast number of (external or internal) stimuli are represented in awareness at any given moment (Posner, 1994). That is, from moment-to-moment, attention gates through which information reaches WM; and WM serves as the temporary storage mechanism of conscious detail (Bor & Seth, 2012). In this way, internal attention determines the focus of awareness and IDC at each moment; and, by doing so repeatedly, also sustains or shields IDC by suppressing both external source selection (Smallwood, 2013a) and other internal distractors. Accordingly, the capacity for crucial mental tasks, for example, planning an important event in the future, requires the ability to (at least temporarily) filter and gate external and internal distractors as well as to maintain and update WM with the (event) relevant information. As such, the inability to filter out *external* distractors (e.g., a noisy train station) results in impaired capacity for focusing on important internal goals and processing. This is known as the perceptual decoupling hypothesis—that the inhibition of perceptual inputs into WM is necessary for the temporal continuity of IDC (Smallwood, 2013a; Smallwood & Schooler, 2015). Likewise, the inability to filter and suppress internal distractors, for example, associative memories and thoughts, similar results in the inability to sustain contextually important IDC or attention to external stimuli.

To account for findings that mindwandering and sustained (external) attention are mutually inhibiting processes, and the corresponding decoupling hypothesis, Smallwood (2013a) proposed the process-occurrence framework—a seminal conceptual model of mindwandering. First, the process-occurrence framework argues that both mindwandering and on-task (external) attention are subserved by similar (shared) cognitive processes—i.e., both phenomena (i.e., higher-order processes) emerge from a general cognitive architecture (i.e., interacting lower-order processes). Thus, because some cognitive processes resources are domain-general (Baddeley, 2012; Teasdale et al., 1995), if mindwandering occurs, it reduces available shared resources for simultaneous attention to on-going tasks. Accordingly, the process-occurrence framework highlights the importance of such a general cognitive architecture in IDC, and critically, the importance of distinguishing between the higher-order processes/phenomena (i.e., mindwandering, on-task attention) that emerge from the lower-order processes of the cognitive architecture that subserve these phenomena. We similarly conceptualize IDC as a higher-order process emerging from lower-level processes. This conceptualization is in-line with multi-component cognitive (McVay & Kane, 2013; Smallwood, 2013a) and neuroscientific (Andrews-Hanna, Smallwood, & Spreng, 2014; Axelrod et al., 2017) accounts of IDC. While different accounts of IDC implicate (or at least empirically test) different components (e.g., attention, executive control, self-referential processing, scene construction; Axelrod et al., 2017; Smallwood, 2013a), these various accounts commonly conceptualize IDC as emerging from the interaction of lower-level processes, components, or neural substrates.

A Dynamic and Complex System

Accordingly, our understanding of the specific mechanism(s) through which internal attention shapes (mal)adaptive IDC is initial and modest (Andrews-Hanna et al., 2014; Axelrod et al., 2017; Christoff et al., 2016; Chun & Johnson, 2011; McVay & Kane, 2010; Smallwood, 2013a; Smallwood & Schooler, 2006). We argue that our capacity to predict and model (mal)adaptive IDC may be significantly facilitated through understanding the *complexity and dynamics of how internal attention interacts with other cognitive processes* that together form a system that subserves higher-level process of (mal)adaptive IDC.

First, **complexity in this system** arises from the multiple inter- and trans-actions between low-level cognitive processes (Smallwood, 2013a) including internal attentional selection, working memory, mnemonic and emotional factors as well as contextual demands on attention (Axelrod et al., 2017; Dixon et al., 2014). Indeed, the interaction between these processes is characterized by a circular causality (Kelso, 1995; Lewis, 2005). For example, internal and external attention gate the information entering WM, yet the contents or activated representations in WM also guide internal and external attentional selection (Hollingworth & Luck, 2009; Nobre & Stokes, 2019). Similarly, attention amplifies perceived emotional experience (similar to how external attention increases perceptual vividness; Mrkva, Westfall, & Van Boven, 2019). In turn, emotional states amplify attentional selection by facilitating the processing of- or disrupting on-going tasks and reorienting to- emotionally relevant information (Öhman, Flykt, & Esteves, 2001; Smallwood, Fitzgerald, Miles, & Phillips, 2009; Vuilleumier, Armony, & Dolan, 2003).

Second, not only do these lower-order processes have complex inter- and trans- actions in any given moment, but these complex interactions unfold dynamically from moment-to-moment in time. A **complex dynamic system** emerges from these circular causalities in time: The states of each process in the current moment are determined by the systematic interactions between those states in the previous moment(s) which then influence the state of each process in the next moment(s)—repeatedly over time (see Figure 1B).

Furthermore, this complexity in time may also subserve **dynamics of higher-level (mal)adaptive internally-directed cognitive processes**—most notably, the temporal stability or variability of thought content. For example, while goal-directed thinking and mindwandering are both expressions of IDC (Christoff et al., 2016), they are characteristically different in terms of stability/variability of thought content in time. Goal-directed thinking is characterized by constrained internal attention to goal/task-relevant information in a purposeful, organized manner, whereas mindwandering is characterized by goal/task-irrelevant fluctuations in thought content that is processed in a more spontaneous, less deliberate way. Temporal epochs of underlying (i.e., lower-order)

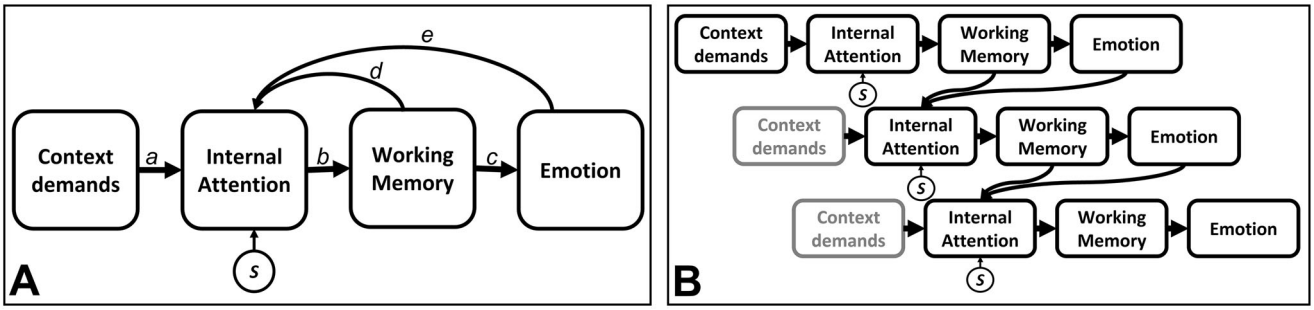

Figure 1. Overview of the model. *S*: stochastic/random effects on selection. (A) Illustrates the *momentary state in context* of the model. Path *a* reflects cognitive control, path *c* reflects emotional reactivity and path *e* reflects cognitive reactivity. (B) Illustrates how the model unfolds in time.

narrowed the scope of internal attention (e.g., when selection is constrained to representations associated with content in WM) subserve periods of temporal *stability* in thought content (Whitmer & Gotlib, 2013). For example, when a thought or memory leads to a thematically related thought or memory, conscious (stream of) thought may be experienced as coherent (Epstein, 2000). Likewise, moments in which internal selection is not constrained (or fails to constrain) to information associated with contents in awareness (or WM) are more likely to lead to a spontaneous shift in the thematic content of thought. Such moments subserve higher-order episodes of temporal *variability* in thought content, such as episodes of task-unrelated thoughts (when there is an on-going task to attend to) or mindwandering (when there is no urgent task to attend to) (Axelrod et al., 2017; Christoff et al., 2016; Ellamil et al., 2016; Engen & Anderson, 2018).

A Dynamic and Computational Systems Framework to Model Internal Attention in Internally-Directed Cognition

We propose that a dynamic systems framework may help advance our understanding of the nature and function of internal attention in (mal)adaptive IDC. We use the term dynamic systems to refer to the broad concept of a multi-component system that are characterized by change over time (Juarrero, 2000; Lewis, 2005) increasingly used in psychological science (Fried & Cramer, 2017), rather than significantly more complex dynamic systems *theory* common in mathematics or physics (Thurner, Hanel, & Klimek, 2018). First, dynamic systems entail processes (or "components") that can change over time that have multiple, reciprocal influences on one-another, are recursive (i.e., the system iteratively interacts and unfolds over time), and are not constrained to linear inter-component relations (e.g., may be subject to dampening or amplifying effects). Second, dynamic systems help to explain how "wholes" (what we refer to as "higher-level processes") emerge from parts (what we refer to as "lower-level" processes) (Lewis, 2005). For example, how does "higher-order" IDC emerge from the interactions of "lower-order" processes, such as internal attention, working-memory, and emotion—as called for by the process-occurrence framework (Axelrod et al., 2017; Christoff et al., 2016; Smallwood, 2013a)? By adapting a dynamic systems framework and terminology, and building on the understanding of how the basic low-level processes or components of the system dynamically inter- and trans- act, we may be able to understand how (mal)adaptive higher-level IDC emerges.

Furthermore, we propose that *formal computational modeling* may be particularly instrumental in advancing the understanding of internal attention in (mal)adaptive IDC (Epstein, 2008; Grahek, Schaller, & Tackett, 2021). Moreover, formal computational modeling may be *necessary* for dealing with- and understanding- the complexity of the dynamic system subserving IDC (see Epstein, 2008; Farrell & Lewandowsky, 2010; Lewis, 2005 for reviews). Formal computational models are designed to formalize and computationally implement and quantify theory by translating the key principles of a theory into computational rules and/or equations (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021). Critically, "formal" computational models should not be confused with "data" computational models (Borsboom et al., 2021). Data models are designed to process, test, and understand *empirical* data (e.g., SEM, network models, reinforcement learning models, evidence accumulation models (Gershman, 2016; Haslbeck, Ryan, Robinaugh, Waldorp, & Borsboom, 2019; Heathcote, Brown, & Wagenmakers, 2015). Our focus here is explicitly on the former—formal computational models.

First, the development of computational models—the theory formalization process—reciprocally and iteratively facilitates the development of the theory behind the model by necessitating that all or most assumptions or key parameters of a theory are explicit (Borsboom et al., 2021; Epstein, 2008; Grahek et al., 2021; e.g., are causal relations between system components (non)linear?). Second, by permitting mathematical manipulations of key parameters and their inter-actions in the dynamic system, computational simulations can provide a novel and more precise understanding of, and predictions about, the phenomena of interest than conceptual theory alone (Huys, Maia, & Frank, 2016). Such benefits include the capacity to run computational simulations of the model and thereby quantify and visualize momentary states as well as temporal trajectories of the dynamic system under various theoretical scenarios/parametric conditions (e.g., context, initial states, or parameters). In turn, computational modeling enables formal simulations testing how complex systems unfold in time (Borsboom et al., 2021; Epstein, 2008). Critically, such simulations allow

for exploring whether a proposed formal theory (e.g., A2T) provides a *plausible explanation* of the phenomena of interest [i.e., primary explanada (van Rooij & Baggio, 2021); e.g., repetitive negative thinking]. Finally, formalized models permit the comparison of simulated and observed experimental effects and results (i.e., secondary explanada; see Borsboom et al., 2021; van Rooij & Baggio, 2021). This facilitates the examination of the explanatory scope and empirical verisimilitude of the model and theory (Borsboom et al., 2021; Epstein, 2008; van Vugt & van der Velde, 2018).

Attention-to-Thoughts Model

"All entities move and nothing remains still."—Heraclitus.

Thus, we propose the A2T model—a conceptual and computational dynamic systems model of internal attention in IDC. Through the model we aim to, first, conceptually and computationally define *momentary states* of this dynamic system of attention, thought, and emotion; and second, we aim to simulate and predict differential *trajectories* of this dynamic system. To do so, the A2T model characterizes how internal attention affects and iteratively transacts with working-memory and affects processes, in time, as a function of contextual demands for sustained focused attention.

The model thus focuses on two explanatory levels. First, **momentary states in context**—how the components of the model causally inter- and trans-act in each moment in time (temporal resolution: seconds). The results of these inter- and trans-actions in time determine each of the following states of the system components: (1) the current internal attention selection likelihood or "bias" to attend to specific representations (e.g., negative *vs.* neutral representations), (2) the current representations active in WM, and (3) the current affective state (e.g., degree of momentary negative affect). Critically, each *momentary state* (Time t) is influenced by the previous (Time $t-1$) and affects the subsequent (Time $t+1$), momentary state. That is, each momentary state is causally "linked" to its previous and subsequent states, creating a temporal "chain" or vector of momentary states of the system. Accordingly, the second level of explanation focuses on temporal dynamics or **trajectories**—how momentary states of the dynamic system unfold in time and content, from moment-to-moment (temporal resolution: minutes, hours). Critically, the likelihood and frequency of specific trajectories constitute individual differences in key aspects of higher-level processes (e.g., mindwandering, repetitive thinking, cognitive dyscontrol) that emerge from the moment-to-moment unfolding of momentary states or lower-level processes.

Utilizing A2T, we sought to characterize each of the following: (1) how internal attention selection is expressed from moment-to-moment; (2) how internal attention drives and iteratively transacts with WM, affective states, and how such components of the dynamic system feedback onto- and thereby influence- internal attentional selection of thought content; (3) how high and low contextual demands for focused attention to task-relevant information may moderate relations of internal attention, WM and affect; and (4) how individual differences in the magnitude of influence—between WM, affect and internal attention—impact emergent trajectories of the system. Furthermore, by addressing these questions using the A2T model, we demonstrate how dynamic temporal patterns of the system may subserve higher-level (mal)adaptive IDC. Finally, we use A2T to conduct a series of experimental simulations to test and illustrate how the dynamic internal attentional system, under different contextual demands and individual difference parametric conditions, may subserve (mal)adaptative IDC and thereby may contribute to metal health.

Model Synopsis

First, the proposed A2T model characterizes the dynamic, inter-, and trans-action, from moment-to-moment in time, of four key lower-order cognitive-affective components. These components constitute the repeated momentary process of selectively attending to internal representations in working-memory, such as verbal thoughts and memories (Alderson-Day & Fernyhough, 2015; Andrews-Hanna et al., 2014; Dixon et al., 2014; Kanske, Plitschka, & Kotz, 2011), and their influence on the affective state (LeDoux & Brown, 2017). The model components are (1) *Contextual Demand*: Degree that context demands goal-directed focused attention ranging from high-demand task-oriented states to low-demand mind-wandering states; (2) *Internal Attentional Selection*: Selection of internal mental representation(s) for preferential processing and thereby access to working memory; (3) *Representations in Working Memory*: Representations momentarily active in working memory; and (4) *Affect*: The momentary subjective hedonic tone of experience (degree of positive and negative affect). Each component has a distinct function and process (see Table 2). Yet, from their myriad constellations of simple interactions over time (Kelso, 1995) emerge the various complex (higher-order) expressions of IDC (Smallwood, 2013b).

Model Components

First, *internal attentional selection* is at the functional center of the proposed dynamic system subserving IDC and accordingly at the center of A2T model. At each moment in time, various sources of internal information, such as stimulus-driven spontaneous associative thoughts and memories as well as goal-directed thoughts, compete for processing.[1] The selection determines what internal information receives

[1]Computational approaches have helped to advance theories of (external) attention, and in cases where different theories converge to similar predictions, to then clarify theoretical differences in underlying processes and computations (e.g., percept *vs.* exemplar similarity estimation or signal-over-noise detection) (see Logan, 2004 for a review). Such theories are based on the rich history of psychophysics research and accordingly focus on external attention (Chun et al., 2011). A2T focuses on internal attention *in IDC*, and integrates other components into a dynamic system subserving IDC, rather than focus solely on attentional processes (Smith & Sewell, 2013). Previous theories nevertheless influenced conceptualizations in A2T—most notably executive control driven attentional weights for biasing and optimizing selection, and the inclusion of noise/stochastic processes (Logan, 2004).

Table 2. Definitions of A2T key terms and model components.

Term	Category	Scale	Definition
Context	System component	Seconds	Degree that current context demands goal-directed focused attention.
Internal attentional selection	System component	Seconds	Momentary probability of internal selection given previous state
Representations in WM	System component	Seconds	Representations momentarily stored and accessible in working memory
Affect	System component	Seconds	The momentary the subjectively experienced hedonic tone (degree of positive and negative affect)
Momentary state in context	Emergent characteristic	Seconds	Momentary constellation of the components of model—i.e., momentary context, momentary affective state, momentary representations stored in working memory)
Trajectory	Emergent characteristic	Minute(s)	States in context in time
Vulnerability	Emergent characteristic	Days, months, years	Chronicity of trajectory or states in time, over longer time epics and across contexts

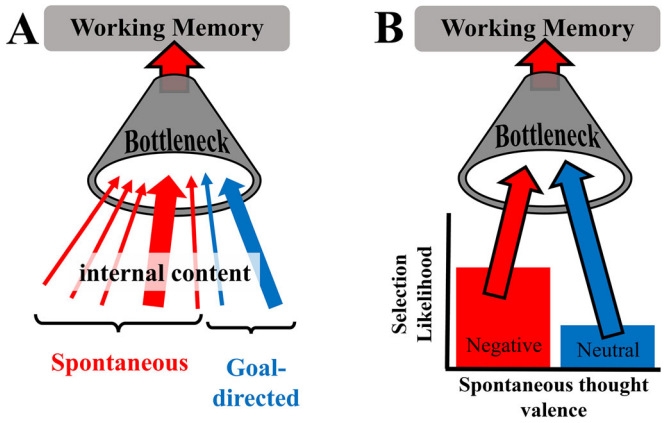

Figure 2. Illustration of selection of internal content. (A) Visualizes the competition by several objects. (B) Visualizes how competition out-come can be represented as likelihood distributions.

processing priority from the limited capacity in working memory (see Figure 2A). Selection can be based on any number of dimensions reflecting the "features" of thought content (e.g., degrees of emotionally-associated valence, self-referentiality, temporal orientation, etc.; see Andrews-Hanna et al., 2014). For simplicity in outlining and illustrating the model, in this paper, we focus on (degree of) negative valence as the key dimension on which selection is based.[2] The momentary state of internal attentional selection is represented, and may be formalized, as a probability distribution biasing selection in favor of certain representations over others in any given moment (see Figure 2B). Critically, in each moment this selection probability distribution changes as a function of the momentary states of the other components (their mechanism of influence is explained below). An additional source of influence on the selection is stochastic/random factors, such as random neural (system) noise (Buzsáki, 2006; Logan, 2004; Shadlen & Newsome, 1994; van Vugt & van der Velde, 2018). This randomness is essential to ensure a degree of *flexibility over time* such that the system does not become "stuck" indefinitely in a certain (biased likelihood) state due to internal feedback loops without an intervening stochastic event (see *Common temporal trajectories* section below).

Second, the contents or ***representations in working memory*** are the momentary outcomes of internal selection. These include mental experiences, such as verbal thoughts (internal speech), internal imagery, and associative memories (Alderson-Day & Fernyhough, 2015; Oberauer & Hein, 2012). These representations are stored in- and as such subject to the constraints of- short-term memory and accordingly "exist" for short periods of time (until decaying or reattended/selected) (Baddeley, 2012).[3] During these time periods, these representations have brief (i.e., phasic) influence over likelihoods of selection (internal attentional selection component) by biasing likelihoods in favor of content-congruent internal representations (Hollingworth & Luck, 2009; Smallwood, 2013a), similar to the biasing effects of selection history on external attention (Awh et al., 2012) (see Figure 1A).

Third, the **affect** component reflects the consequence(s) of the representation in working memory on the current negative affective state (i.e., the more negative representations in WM trigger more negative affect; Engen & Anderson, 2018; Ruby, Smallwood, Engen, & Singer, 2013; Siemer, 2005). Like the ***representations in working memory*** component of the model, affect has phasic influence over the likelihoods of selection by biasing selection in favor of affect-congruent internal representations. This is assumed to occur through spreading activation along a shared associative network between the current affect tone and long-term memory representations (e.g., negative affect increases activation levels of memories associated with negative affective experience) (Eich, 1995; Williams et al., 2007) (Figure 3).

The fourth component, *contextual demands for focused attention*, reflects the degree to which a person, at each moment in time, is under (external or internal) demands to selectively attend to specific information (Buetti & Lleras, 2016; Lavie, 1995; Smallwood, 2013a; see Figure 4). For example, external contextual high-demands for focused attention may be participating in a cognitively demanding

[2]However, the features/dimension on which selection is based can be changed when applying the model to specific topics of interest.

[3]For computational/simulation purposes we view this component as a repeatedly updating first-in-first-out storage device containing the last five selected representations/objects.

experiment; an internal high-demand may occur when a person tries to recall important details from a specific autobiographical memory. We thus refer to high-demand contexts as task-oriented states and low-demand contexts as mindwandering states (Christoff et al., 2016). Within the model, greater demand is translated into greater constraints on the momentary attentional selection bias to negative representations (i.e., more contextual demands reduce internal selection bias driven by the current emotional state of the system) (Christoff et al., 2016; Whitmer & Gotlib, 2013). The less contextual demands, i.e., more mindwandering state, the less bias selection in favor of particular representations.

Accordingly, in the model, the contextual demand component has two (computational) functions. First, contextual demand delineates what information is contextually- (i.e., task-) relevant (Buetti & Lleras, 2016). This is similar to the functions, in external attention, of top-down signaling of features or task sets (Corbetta & Shulman, 2002; Logan, 2004). For example, contextually-relevant information may be recalling a shopping list (wherein information is emotionally-neutral). Alternately, imagining an upcoming confrontation with an unpleasant coworker (wherein certain context-relevant information is emotionally evocative). Second, contextual demand additionally functions to indirectly moderate (attenuate) the influences of representations in WM and affect components of internal attentional selection. The degree of attenuation increases as greater contextual demands for focused attention increase.

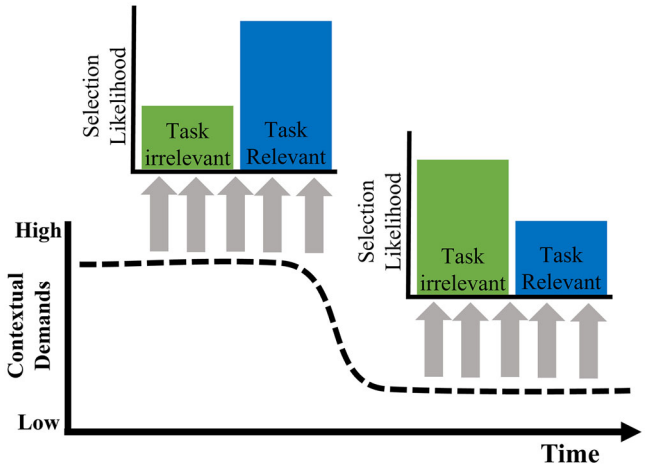

Figure 3. Illustration of the effects of contextual demands for focused attention on selection of task relevant and irrelevant information.

Model Paths: Feed-Forward and Feed-Back

The components of the system are linked through several feedforward (see Figure 1 paths *a*, *b*, and *c*) and feedback paths (*d* and *e*). For simplicity, the strengths of causal links are assumed to be constant between people, except for two

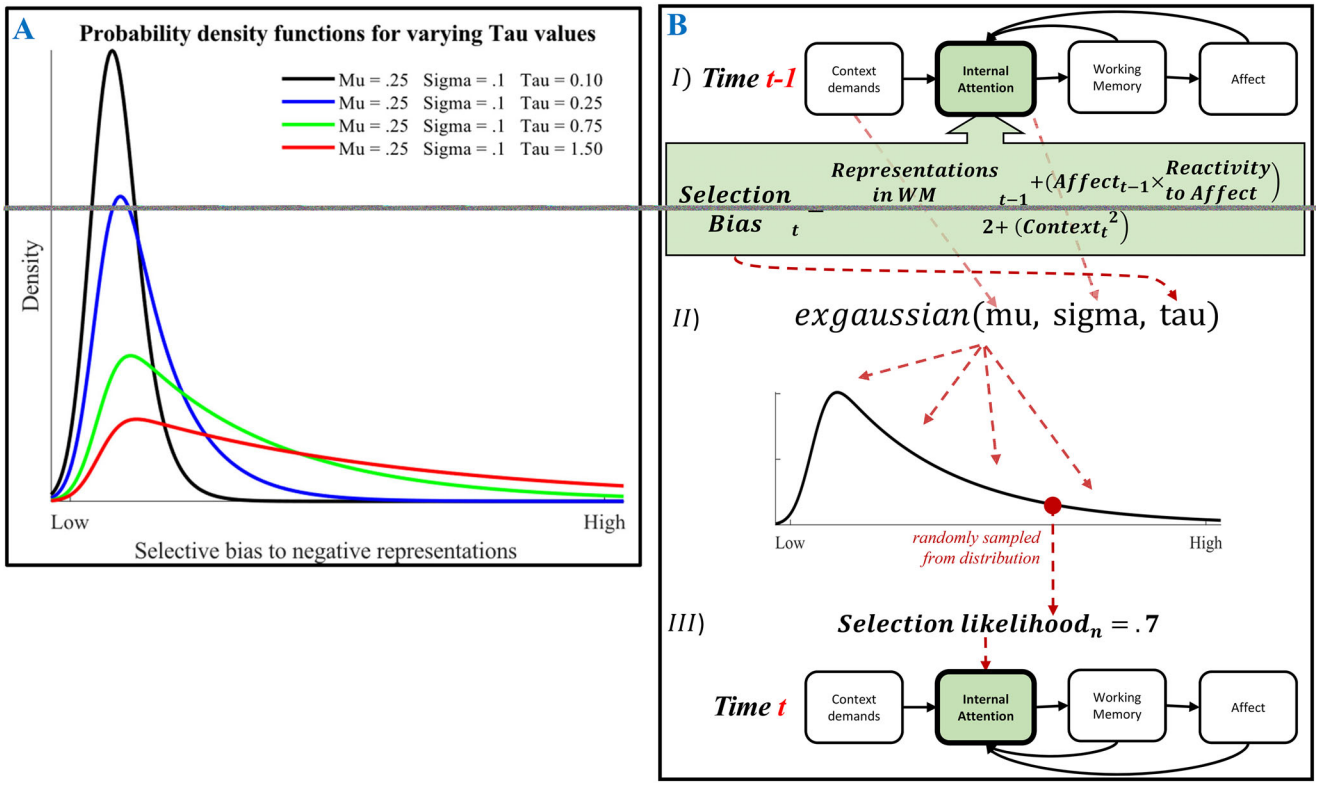

Figure 4. (A) Illustration of change of ex-gaussian distribution as a function of tau values. As tau increases (i.e., exponential rate), the more the distribution is skewed to higher levels of selective bias to negative representations. (B) Illustration of computational procedure of determining tau/selection bias at *Time t*. In brief, first the states of contextual demands (*Time t*), representations in WM (*Time t-1*) and affect (*Time t-1*) are integrated using formula to determine the tau parameter (part I) together with the mean and variance (part II). From this distribution function a single value is randomly sampled (part III). This likelihood value determines the selection likelihood of attending to a negative representation at *Time t*. Finally, in simulations of the model, to constrain values between 0 and 1, this selection likelihood is converted with tangent hyperbolic function with 0 imputed for values <0. Based on this selection likelihood value an "uneven coin-toss" is simulated to decide whether a negative or neutral representation is selected/entered into WM.

important paths we believe are crucial for accounting for individual differences—(affective) *reactivity to representations in WM* (path *c*) and (attentional) *reactivity to affective state* (path *e*). These paths are crucial to understanding why certain trajectories are likely to occur more frequently in certain individuals relative to others (see *Common temporal trajectories* below). Accordingly, the paths of reactivity to representations in WM (path *c*) and reactivity to affective state (path *e*) are the only two elements in the model that may vary between individuals in their relative causal strength. As such, the causal strength of these paths or processes is not predicted by the model, rather, between-person variability in these paths are moderated by factors extraneous to the system (Fried & Cramer, 2017). As such, individual differences in these paths are crucial for generalizing and applying the model to fields of research in which core questions and interests deal with between-person variability—for example, individual differences in cognitive vulnerabilities to depression and anxiety (see *Higher-level processes: cognitive vulnerabilities* below).

First, *affective reactivity to representations in WM* (path *c*) or "emotional reactivity" refers to the change in an emotional state and its intensity (peak amplitude) in response to emotional information (Davidson, 1998). Reactivity to representations in WM is an important individual differences factor in mental health, as persons high on repetitive thinking, depression, or anxiety show increased negative emotional response when instructed to direct attention to negative *internal* experiences (mostly through experimental manipulations, such as "Think how you feel inside."; Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008; Watkins, 2008). In the proposed A2T model, reactivity to representations in WM is reflected in the effect of cognitive representations in WM on affect (see Path *c* in Figure 1A). Accordingly, when an emotionally associated representation is selected into WM it activates a corresponding affect state (affective/emotional reaction). The "higher" an individual is on reactivity to representations in WM, the stronger the negative representations in WM trigger a momentary negative affect.

Second, *cognitive reactivity to affective state* (path *e*) or "cognitive reactivity," refers to the magnitude of increase in the probability of negative cognitions (thoughts) in response to negative affect (Scher, Ingram, & Segal, 2005). For example, in both experimental and naturalistic settings, depressed mood has been shown to increase the likelihood of recalling negative memories (i.e., mood-congruent memory) (Teasdale & Barnard, 1993). In the model, cognitive reactivity to affective state is reflected in the influence of affect on the internal attention selection likelihoods in favor of affect-congruent representations (Figure 1A Path *e*). For example, the greater momentary negative affect, the greater internal attentional selection will be biased toward negative representations.

Computational Formulations

Momentary States of the System

As detailed above, *internal attentional selection* is at the functional center of A2T. Momentary states of each component converge onto internal attentional selection—determining the likelihood of attending to a negative representation state of the internal attention component at Time *t*. To reiterate, here we illustrate the role of internal attention in IDC using an A2T model which focuses on attention to negative *vs.* neutral thoughts. Accordingly, selection likelihoods are based on the "feature" or dimension of (degree of) negatively valenced representations. Notably, the model can be applied to other (potentially multiple) features/dimensions of interest (e.g., self-referentiality, chronic goal relevance).

Computationally, the likelihood of attending to negative representations is drawn from an ex-Gaussian distribution whose parameters (mean, variance, and exponential rate; see Table 3) change, from moment-to-moment, as a function of the state of different system components. The *mean* parameter is determined by information from contextual demands of what is task-relevant information, and more specifically, the degree (from 0 to 1) to which emotionally negative or neutral information is task-relevant. For example, the mean parameter will be .5 when modeling an individual trying to recall a list with an equal number (50/50%) of emotionally neutral and negative words. The *variance* parameter represents the stochastic element of internal attentional selection, such that the greater the stochastic element, the less the selection likelihood distribution constrains around the mean. Finally, the *exponential rate* (tau) parameter reflects the degree of momentary selection bias to negative representations. Rate is determined by the momentary interaction of representations in WM, affect, and contextual demands (see Equation 1 below). Essentially, as the rate (i.e., negative bias) increases, the distribution accordingly biases/shifts toward an overall greater likelihood of attending to negative representations (see Figure 4).

Momentary States: Selection Bias (Exponential Rate/Tau).

The mean and variance parameters reflect typically stable factors (e.g., context is usually stable over short periods of time) and, as such, do not vary from moment-to-moment. In contrast, the rate parameter, representing selection bias, is most subject to temporal changes because it is determined by the momentary states of the system components. As such, the rate parameter is central to understanding how the model trajectories unfold over time. This selection bias (exponential rate, tau) can be represented as a heuristic formula (Equation 1):

$$\text{Selection bias}_{i,\ t} = tau = \frac{\text{Representations in WM}_{i,t-1} + (Affect_{t-1} \times \text{Reactivity to Affect}_{i})}{2 + (Context_{i,t}{}^{2})} \quad (1)$$

Table 3. Parameters for the ex-Gaussian distribution of likelihood of attending to negative representations.

Parameter	Meaning
Mean (mu; μ)	Proportion to which negative information is context-/task-relevant. The more the current task/context requires prioritizing negative representations, the distribution center accordingly shifts to "higher" selective bias.
Variance (sigma; σ)	Degree of stochastic element of internal attention. As the stochastic element increases, the wider the distribution becomes, and so selection likelihood randomly varies more from moment-to-moment.
Exponential rate (tau; τ)	Degree of bias to negative information. Integrates the states of representations in WM and affect (Time $t-1$) and contextual demands (Time t). As tau increases, the distribution shifts more toward higher negatively biased selection likelihoods.

where i represents the individual person/model, t the specific moment in time. Accordingly, *selection bias* is a value between 0 (no bias to negative stimuli) to 5 (complete bias to negative representations) and is determined by values of the following elements: (a) the product of $affect_{t-1}$ and *reactivity to affective state*, and (b) the sum valence of *representations in* WM_{t-1}.[4] The average of these two elements is divided by (c) $Context_t$, the demand for focused attention at that moment in time. *Context* is squared and placed in the denominator to reflect the strong constraining influence of contextual demands on internal attentional processes (i.e., the higher the contextual demands for focused attention the less influence for non-context-related biasing factors, such as affect).

Momentary States: Affect. The degree of momentary subjective negative *affect* is determined by the affective valence of representations active in working memory (which is determined by internal attentional selection) and reactivity to representations in WM. This can also be represented in a heuristic formula (Equation 2):

$$Affect_{i,t-1} = \frac{Reactivity\ to}{representations\ in\ WM_i} \times \frac{WM}{represention_{i,t-1}} \quad (2)$$

Momentary States: Representation Selection. After the selection likelihood distribution has been determined by the parameters (mean, variance, tau) a single value at *Time t* (*Selection likelihood*$_t$) is sampled from the ex-Gaussian selection likelihood distribution. Finally, in simulations of the model, to constrain values between 0 and 1, this value is converted using the tangent hyperbolic function with 0 imputed for values < 0. Based on this final selection likelihood$_t$ value an "uneven coin-toss" is simulated to decide whether a negative or neutral representation is selected/entered into WM. This simulates how, from any number of different representations competing for selection into WM, a single representation is selected, and this selection outcome is influenced by the momentary bias of the internal attentional selection component.

[4]As described in *Model components* section, the Representations in WM component is computationally implemented a short-term storage of five representations. Where neutral representations have a value of 0 and negative representations a value of 1.

Inter- and Trans-Action of Model Components in Time

These heuristic equations help to illustrate how each contextualized momentary state of the system (a) converges on internal attention and thereby a Selection likelihood value at time t, (b) how this selection likelihood is influenced by the moment state values of *representations in WM* and *affect* at *Time* $t-1$, and (c) how *representations in WM* and *affect* (at *Time t*) are influenced by the probabilistic outcome of selection likelihoods of the *internal attention* component at *Time t*. This formalization of model components' inter- and trans-actions, from moment-to-moment in time, facilitates modeling of the system's temporal dynamics and complexity (see Figure 1B). For example, while affect at *Time t* is determined by the representations in WM at that moment in time, affect also influences internal attention at *Time* $t+1$ which, in turn, will determine representations in WM and thereby affect at *Time* $t+1$ and so on. Accordingly, to understand the current state of the system, at any moment in time, it is essential to understand the preceding state(s). Consequently, by computationally linking together a series of preceding and subsequent states (i.e., momentary state in context at Time *1* to *t*), a dynamic trajectory (i.e., a "greater whole") emerges representing the unfolding of attention, thought, and affect over time, as detailed below.

Temporal Trajectories from Sequential Momentary States

A central function of the A2T model is its capacity to characterize and make predictions about how one momentary state in context leads to the (likely) next momentary state, and so on, emerging over time as a *trajectory*. As illustrated above, the dynamic trajectories of the system are a function of the moment-to-moment values of each model element, and critically, the inter- and trans-actional feedback (paths *d* and *e*) between the components and inter-component paths within the system in time. Accordingly, one constellation of initial values of the components is more likely to lead to a specific trajectory than other constellations of initial values. For example, an initial state of low contextual demands, concurrently with a negative affect state, is more likely to lead to a trajectory of repeated negative content in WM and negative affect (i.e., repetitive negative thinking) than would an initial state of high contextual demands with a neutral affect state. Thus, from different initial momentary states in context (i.e., momentary constellations of the components within the system), we are also able to characterize more and less adaptive cognitive process trajectories.

For example, trajectories may be characterized by greater stability or, conversely, variability in thought content (i.e., content of representations in WM and related on-/off- task though) and related fluctuations in affect. Stability is reflected in temporal sequences of thematically similar/related thought content (e.g., a series of task-related thoughts). Temporal variability reflects the opposite—sequences of thoughts that are thematically unrelated to each other (e.g., frequent shifts in thought content, such as between task-related/unrelated thoughts). The degree to which stability or variability is (mal)adaptive is context-dependent (Watkins, 2008). For example, in some situations (e.g., a chess game), stability may be more adaptive. In other contexts (e.g., when trying to brain-storm), variability may be more adaptive. However, trajectories of high stability or variability may also be maladaptive, especially when such trajectories become chronic or so habituated that they are insensitive to contextual demand (Christoff et al., 2016).

Simulations of Attention-to-Thoughts: Modeling Dynamics of Internally-Directed Cognition

We conducted a series of simulation experiments using A2T. Broadly, we sought to explore how temporal trajectories emerge from momentary states of the lower-order processes as a function of changes in contextual demands for focused attention as well as the valence of context-relevant information (i.e., to what degree is task-relevant information also negatively valenced). Such simulations may be useful in several ways: to better understand (and visualize) how temporal trajectories of the dynamic system emerge and unfold; to better characterize the momentary and temporal inter- and trans-action of system processes (e.g., attention and affect); to examine whether proposed A2T model formalization sufficiently characterizes all or most important assumptions or parameters of the theory and, as needed, to refine the A2T model; as well to examine the explanatory scope of the theory and model by contrasting simulation findings to observed empirical phenomena and specific experimental findings—and thereby an empirical validation process of the theory and formal model (Borsboom et al., 2021).

Simulation Procedure

In each simulation, we compute the values of each of the four model components, in each moment in time, based on the values of the other components at that- and previous-moments in time (see Figures 1 and 5 and formulas above). Because we are focused on the temporal dynamics of IDC, and the momentary states of the cognitive-affective processes that sub-serve these phenomena, simulations were run over a 5-min time period with a resolution (i.e., discrete time/data points) of 1 s. In each simulation the initial starting points (Time *0*) are identical and we manipulate only the starting values of the relevant factors/components according to that simulation's goal. For example, when examining the role of degree of contextual demands for focused attention, we accordingly run two simulations—one with high demand for sustained focused attention and one in which demand is low. The simulation outcomes are longitudinal data—moment-to-moment values of internal attentional selection likelihoods, representations in WM (i.e., thought content in awareness), degree of negative affect, as well as the temporal features (stability, variability) of the trajectories in these processes over the simulated 5-min period. See Figure 5 below for a visual illustration and details on the simulation algorithm and procedure. The simulation code in MATLAB is openly available at: https://github.com/iftachamir/AttentionToThoughts.

The Effects of High Contextual Demand for Sustained Focused Attention on Temporal Trajectories

High Contextual Demand to Focus on Neutral Information: Sustaining Focused Attention on Neutral Information

The ability to focus attention on an on-going task is crucial for adaptive functioning—from mundane tasks, such as planning dinner to recalling first aid procedures during an emergency. Such contexts of high-demand for focused attention impose strong constraints on internal attentional selection likelihoods. As demonstrated heuristically in Equation 1, the greater the value of the contextual demand denominator, the less "weight" or influence there is for affect or representations in WM (i.e., Equation 1 numerator) on internal attention selection likelihoods. Accordingly, under high contextual demands for sustained focused attention, a task-oriented state and subsequent trajectory initiate and are maintained until contextual demands for sustained focused attention decrease. For example, trying to recall the author and name of a paper to cite requires that internal attention be constrained to associated information in long-term memory until the required information is retrieved.

Figure 5A illustrates a simulation of how selection bias and negative affect are maintained at a low value when high contextual demands for sustained focused attention on neutral information are high. Moreover, because selection likelihoods are drawn from an ex-gaussian distribution (and thus even when the distribution mean is 0 there is still some part of the distribution above 0) there is still a (low) probability for selection of negative representations. As such, there are time points in which negative representations are randomly selected. Accordingly, when this random selection occurs sequentially (see Figure 5A ~190 s mark) there is an increase in the values of both attentional and affective components. Importantly, however, this rise is temporary as the tonic, sustained, effect of contextual demands eventually out-competes "random" or low probability events (Buetti & Lleras, 2016). This simulation illustrates how the system is indeed able to maintain focused attention—i.e., suppress emotionally-evocative but task-irrelevant thoughts—when the contextual demands are high. Finally, these simulation findings (see also below *The effects of low contextual demands*) are in-line with empirical findings showing that subjective reports of mindwandering decrease as a function of task difficulty (Seli, Konishi, Risko, & Smilek, 2018) as

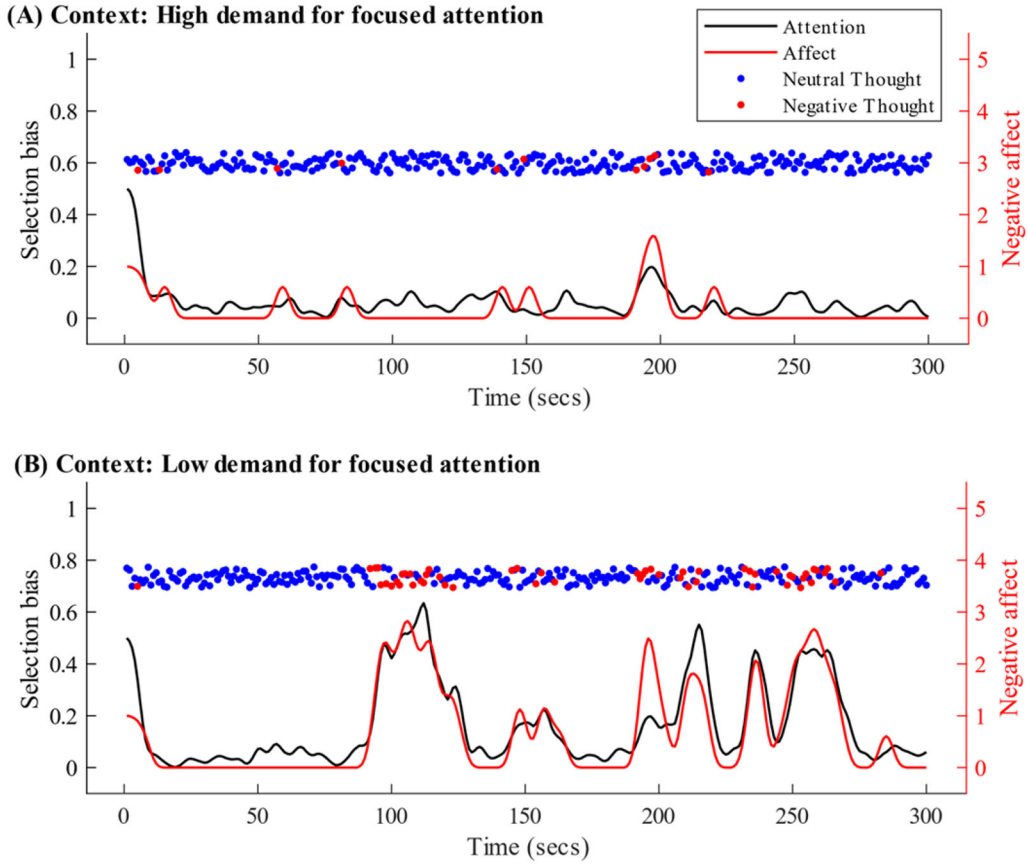

Figure 5. Simulated values in time (5 min) of Internal Attention and Affect components. Red line = degree of negative affect at each moment in time. Black line- = selection bias—the likelihood of selectively attending to negative representations. Importantly, the selection bias is a *likelihood but not necessarily outcome*—i.e., even under high bias there is a small likelihood for neutral representation to be selected. Accordingly, dots represent the valence of the *outcome* of selection in each time point. Red dot = Selected *negative* representation into WM. Blue dot = Selected *neutral* representation. In (A), the contextual demands is set to high (= 4) and in (B) to low (= 2). In both simulations reactivity to- affect and to- representations in WM (= −.25), as well as initial attention (= .5) and affect are set to moderate (= 1).

well as with corresponding Attentional Resources (Smallwood & Schooler, 2006) and Control Failures (McVay & Kane, 2010) hypotheses and findings.

High-Demand Context and Stochastic Nature of Internal Attentional Selection: Random Lapses in Sustained Attention. Lapses in sustained attention, even for people with high levels of attentional control, occur (infrequently) under demands for focused attention. In the A2T characterization of IDC, lapses in sustained attention are expressed as occurrences wherein task-unrelated thoughts are selected (over task-related thoughts). Accordingly, in A2T, such lapses occur in part due to the partially stochastic nature of the internal attentional selection component of the model–wherein selection is represented as a probability and so, occasionally, low probability events occur. These infrequent lapses are reflected visually in the random (yet infrequent) selection of negative task-unrelated representations (red dots) in Figure 5A. Consequently, a task-unrelated representation in WM may bias selection in favor of (current WM) content-related but task-unrelated thoughts, increasing the likelihood of mindwandering. Yet, mindwandering is likely to dissipate quickly due to the sustained biasing effect of contextual demands in favor of task-related information (see Figure 5A).

High-Demand Context to Focus on Negatively Valenced Information: Sustaining Attention to Emotionally-Negatively Valenced Information. When task-related information is *negatively-valenced*, for example when trying to recall important information from a traumatic experience/memory or when planning for a stressful event, contextual demands bias selection in-favor of task-relevant information which, in this case, is also negatively valenced. Once the first cycle of the system is initiated, and *negative- and task-relevant* representations are selected into WM, these negative representations in WM (in addition to contextual demands) will bias subsequent selection in favor of negative- as well as task-relevant- representations. Accordingly, such negative representations in WM should also trigger negative affect. This may then further bias subsequent selection in favor of negatively valenced—affect-congruent—information competing for internal selection. When task-relevant information may be inter-mixed with *negative and neutral* information, such as memorizing a list of mixed negative and neutral words or reflecting on negative and neutral thoughts, this ultimately creates a selection conflict between the contextual demand and affect-driven biases. For optimal performance on a task, attention should be equally (non-preferentially) distributed between all task-relevant information (in this example both negative and neutral words). While the contextual demands

bias task-*related* content, the negative affect will also increase the likelihood of negative task-*unrelated* content. Therefore, emotionally-valenced representations in WM, even if task-related, increase the likelihood for off-task thoughts and impair task performance (Welhaf et al., 2020). Figure 6 simulates a situation wherein contextually task-relevant information is 50% negative and 50% neutral. This simulation illustrates how in-line with contextual demands, attentional selection bias (black line) fluctuates around .5 and the contents of working memory (blue and red dots) similarly randomly alternate between negative and neutral.

The Effects of Low Contextual Demand for Sustained Focused Attention on Temporal Trajectories

In the absence of immediate task demands, the human brain continues to show coordinated and orchestrated neural activity (Raichle et al., 2001). These findings, alongside concurrent introspective reports of mind-wandering (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009), show that IDC occurs, and potentially increases, during periods of low contextual demands (Seli et al., 2018). In A2T, contexts of low demand for focused attention do not impose constraints on selection likelihoods. Indeed, when there is no "current task," the context does not bias selection in favor of task-related information. As such, the mind is free to wander (Smallwood & Andrews-Hanna, 2013) including to (task-unrelated) personally-relevant chronic goals or unresolved problems (Klinger, Marchetti, & Koster, 2018). In this context, the representations in WM, affect as well as the stochastic element of internal attention has greater weight in influencing selection. The contents of WM influence selection such that current representations in WM (*Time t*) bias subsequent (*Time* $t+1$) selection in favor of content-congruent information. Over time, this should result in a higher probability for a series of repeated selection of content-related representations into WM. This is in-line with the phenomenological experience of relative stability of streams of thought (Epstein, 2000). However, the stochastic element also impacts selection such that at random occasions *unrelated* thought-content (to that at *Time* $t-1$) is selected—switching the "theme" of the stream of thought (Christoff et al., 2016; James, 1890). Consequently, under low contextual demands for sustained attention on a task and emotionally-neutral representations held in WM, the stream of thought contents in awareness should be relatively stable (in terms of perceived relation between content from moment-to-moment) with largely random intervals between changes in the content of those representations (see Figure 5B for visualized simulation).

Interactions of Low Contextual Demand for Focused Attention and Affect

As described above (see *Model components*), affect also influences the selection of internal content by biasing selection in favor of affect-congruent content. Therefore, when negative affect is triggered (*Time t*), it biases subsequent (*Time* $t+1$) selection in favor of (negative) affect-congruent representations (e.g., negative thoughts about oneself). If a negative representation is selected into WM (*Time* $t+1$), then this may further bias selection of negative affect congruent representations, potentially resulting in a spiral of repetitive negative thinking driven by the feedback from representations in WM and affect into the selection and so on (Whitmer & Gotlib, 2013; see ~90 s timepoint in Figure 5B). Here again, the stochastic element has an important and potentially adaptive function. In this type of low-demand context, a negative thought-affect spiral may only be exited—meaning that the self-sustaining feedback loop may be disrupted/stopped—due to random selection of non-negative thought content which terminates the current cycle of repetitive negative thinking. Alternately a change in contextual demands may also lead to an exit from the self-sustaining state of the system (see *High-demand context* above).

Accounting For Maladaptive Internally-Directed Cognition Using A2T

A2T may have a variety of applications and implications for better understanding and accounting for a variety of forms

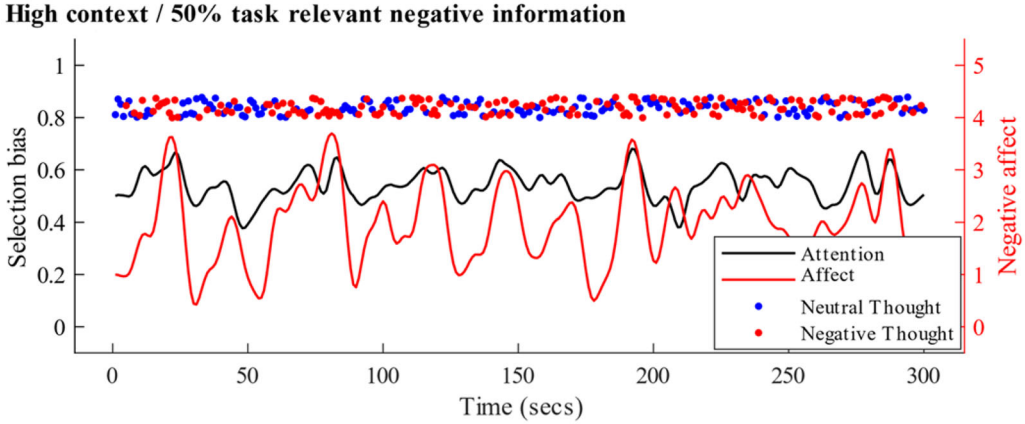

Figure 6. Simulated values when task-relevant information are both neutral and negatively valenced—e.g., recalling a list of mixed negative and neutral words. Relative to Figure 5A, the frequency of negative thoughts (red dots) as well as mean selection bias and negative affect, are higher. All initial component values are equal to those used for Figure 5A.

of IDC, such as dynamics of spontaneous thought, the temporal relations between context, cognition, and emotion, and individual differences in mindwandering and cognitive control (Axelrod et al., 2017; McVay & Kane, 2012; Pessoa, 2008; Schweizer & Dalgleish, 2016). Likewise, and critically, comparing and contrasting A2T model simulation findings to key empirical phenomena and empirical findings provides a powerful process of examination and potential validation of the theory and model (Borsboom et al., 2021; Epstein, 2008). Here, we focus on the application of A2T to account for the emergence of- and individual differences in- maladaptive IDC implicated in the development and maintenance of prevalent forms of mental health problems (Hong & Cheung, 2015; Mansell & McEvoy, 2017; Marchetti, Loeys, Alloy, & Koster, 2016; see Figure 7 for illustration).

Common forms of maladaptive IDC include negative thinking, rumination, and worry (Borkovec, Robinson, Pruzinsky, & DePree, 1983; Ehring & Behar, 2020; Ehring & Watkins, 2008; Nolen-Hoeksema & Morrow, 1991), emotion (dys)regulation (Sheppes, Suri, & Gross, 2015), cognitive fusion and reactivity (Bernstein et al., 2015), cognitive biases, such as interpretation bias (Everaert, Duyck, & Koster, 2014; Everaert et al., 2020) and impairments in cognitive control secondary to negative emotional information in working memory (Grahek, Everaert, Krebs, & Koster, 2018; Joormann & Gotlib, 2008) or dyscontrol over episodic memory (Engen & Anderson, 2018; Hitchcock, Golden, Werner-Seidler, Kuyken, & Dalgleish, 2018). These forms of maladaptive higher-level cognition are characteristically *internal* (focus on thought content and thinking styles), *temporal* (how thought content and style are expressed over time), *affective* (how thoughts invoke affect and vice-versa), and *selective* (e.g., why only specific content becomes repetitive or difficult to disengage from) (Amir, Ruimi, & Bernstein, 2021; Fox et al., 2018; Harvey, Watkins, & Mansell, 2004; Koster, De Lissnyder, Derakshan, & De Raedt, 2011; Nolen-Hoeksema et al., 2008; Siemer, 2005). Accordingly, exploring how the dynamic system of internal attention may subserve these common forms of maladaptive IDC may help illustrate potential applications and implications of the A2T model; and, in turn, provide empirical validation of the theory and model by contrasting simulation findings from A2T to empirical and experimental observations of (mal)adaptive IDC, such as negative repetitive thinking, cognitive (dys)control, and interpretation bias (Borsboom et al., 2021; Grahek et al., 2021).

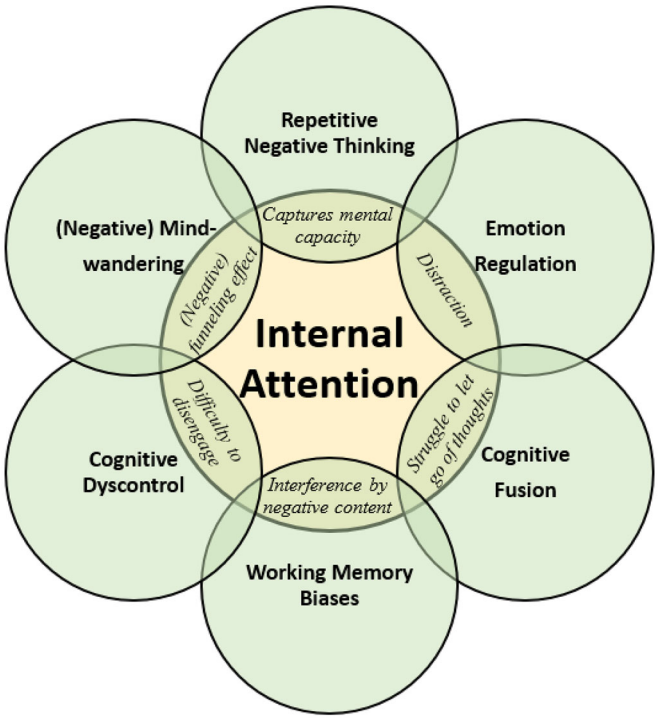

Figure 7. Examples of the role of internal attention in various forms of higher-order (mal)adaptive internally-directed cognition implicated in mental health. The text in the overlap between internal attention (center) and each form of internally-directed cognition (outer circles) delineates the attentionally-driven lower-order processes theorized to subserve each higher-order internally-directed thinking process.

Repetitive Negative Thinking

Repetitive negative thinking has been conceptualized as a style or process of thinking characterized by negative and repetitive thoughts which are experienced as intrusive and difficult to disengage from (Ehring et al., 2011; Ehring & Watkins, 2008). Repetitive negative thinking, as well as related processes of rumination and worry, predicts negative outcomes including momentary negative emotions and, over time, the onset of depression as well as post-traumatic stress and generalized anxiety disorders (Hirsch et al., 2018; Watkins, 2008).

Repetitive Negative Thinking in A2T

In A2T, repetitive negative thinking will be initiated, typically, when a negative representation (thought) is selected into WM in a state characterized by low contextual-demands for sustained focused attention. This representation then triggers or increases negative affect. Consequently, both the representation and affective state bias subsequent selection in favor of content- and affect- congruent thoughts (e.g., "Bad things always happen to me."). This then triggers the same process of selection in favor of negative content and so the same process repeats itself through feedforward (Figure 1A paths *b* and *c*) and feedback (paths *d* and *e*) between internal attentional selection, representations in WM, and affective state. Here, the stochastic element in internal attentional selection is functionally critical. Indeed, the stochastic element enables "random exits" from repetitive negative thinking—selection of non-negative content breaking or attenuating the feedback loop (see Figure 8B time point ~240 s). Without a random element, the model becomes overly rigid, determining that once an RNT state initiates it cannot terminate (except when contextual demands change and focused attention to task-relevant information is required).

Individual Differences in RNT. As described above, the strengths of causal links between contextual demands, representations in WM, and internal attentional selection (paths

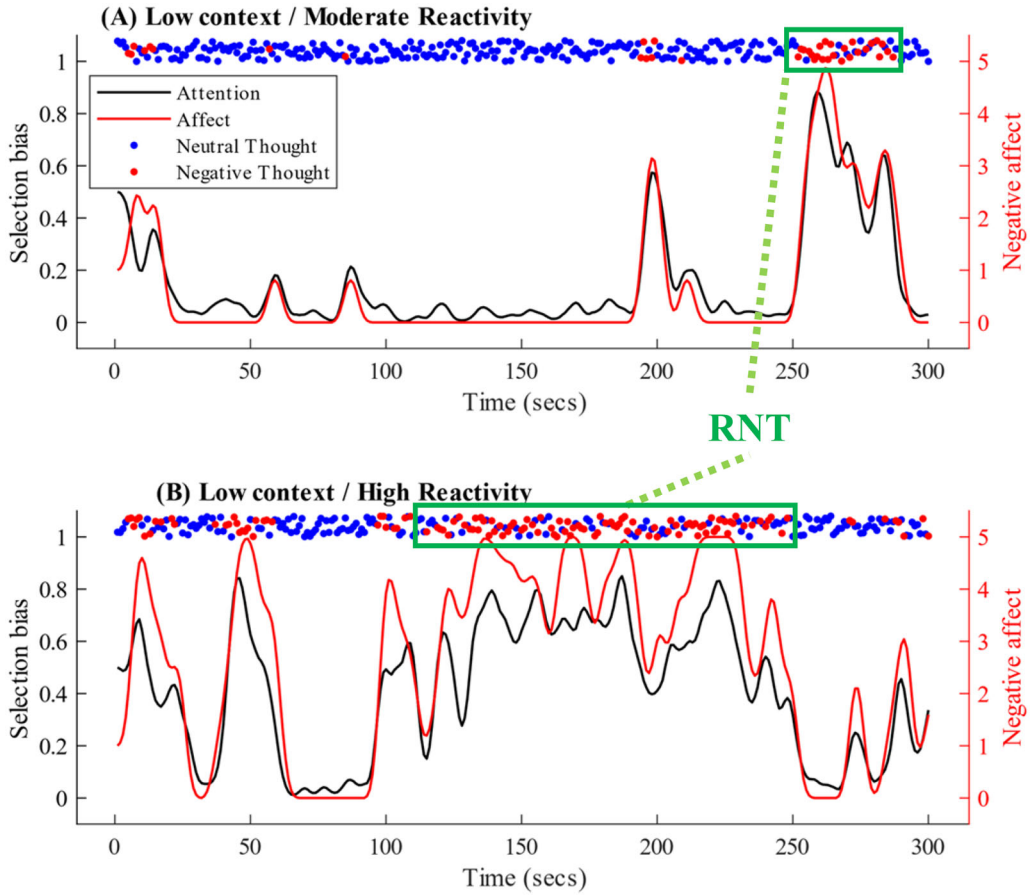

Figure 8. Simulated differences between cognitive and emotional reactivity level (Moderate = 0, High =.25). Contextual demand for sustained focused attention is set to low (= 2). All other initial component values are equal to those used for Figure 5A. RNT = Repetitive negative thinking.

a and *d*; Figure 1A) are assumed to be constant between people. The exception to these are two important paths we believe are crucial for accounting for individual differences, especially when modeling individual differences in maladaptive thinking and maladaptive trajectory likelihoods—(affective) *reactivity to representations in WM* (path *c*) and (attentional) *reactivity to affective state* (path *e*). Critically, these two paths constitute the closed feedback loop from WM onto affect and then back onto the internal attentional selection. Therefore, individuals with higher levels of reactivity have a stronger feedback loop, increasing the likelihood of a self-sustaining state of biased selection of negative representations and heightened negative affect. Figure 8 illustrates simulated differences between models/individuals with moderate (Figure 8A) and high (Figure 8B) reactivity to representations in WM and to affective state. As can be seen, higher reactivity levels increase both (a) the frequency of entering- and (b) the average duration of- states of repeated selection of negative content into WM and related increase in negative affect. Interestingly, research has shown that repetitive negative thinking depletes WM resources (i.e., capacity) and accordingly hypothesized that (internal) attention subserves this effect (Curci, Lanciano, Soleti, & Rimé, 2013; Hayes, Hirsch, & Mathews, 2008). A2T's closed feedback loop between internal attention and WM and affect, and the simulations demonstrating its impact on frequency and duration of repetitive negative thinking may help to explain and account for these effects of repetitive negative thinking on WM capacity.

Summary: Repetitive Negative Thinking in A2T

Critically, A2T predicts that RNT is most likely to emerge under conditions of low contextual demands for focused attention (Seli et al., 2018). A2T also illustrates how the phenomenon of RNT may be occasionally experienced by more or less any person, and not likely a byproduct of malfunction or disorder per se. Furthermore and importantly, A2T also provides a formal theoretical account for individual differences in the frequency and duration of RNT episodes; and predicts that these differences are a function of reactivity to thought content and (negative) affective states (Smallwood et al., 2009). A2T thus provides a dynamic systems account for how the higher-order process of RNT may emerge from the temporal interactions of low-level cognitive-affective processes. The model also explains why, once negative affect is activated, it is difficult to interrupt this process. In general, these predictions are in-line with research showing that RNT both influences- and is influenced by- negative affect (Curci et al., 2013; Poerio, Totterdell, & Miles, 2013; Ruby et al., 2013) and may be expected to function as the transdiagnostic process that cuts across emotional disorders (Ehring & Behar, 2020; Ehring & Watkins, 2008).

Cognitive (Dys)control

Cognitive control refers to the ability to flexibly adapt cognition and behavior to current goals, often used interchangeably with executive control and executive functions (Grahek et al., 2018). Cognitive control accordingly entails various cognitive abilities including inhibiting prepotent or dominant responses and updating and shifting between contents of WM. Cognitive control deficits have been observed in mood disorders and accordingly play central roles in cognitive models of depression and anxiety (Eysenck, Derakshan, Santos, & Calvo, 2007; Grahek et al., 2018). However, it is not clear whether this deficit is specific to control over emotional information or is domain-general (i.e., neutral) (Grahek et al., 2018; Zetsche, Bürkner, & Schulze, 2018).

Cognitive (Dys)control in A2T

Cognitive control and the component of contextual demands for focused attention are closely related. Both serve the same function—prioritizing the processing of task-relevant over irrelevant information (Buetti & Lleras, 2016; Seli et al., 2018). More precisely, we conceptualize cognitive control as a dynamic behavior that emerges from the moment-to-moment competition between the effort to maintain attention on task-related information *vs.* the capturing of attention by salient but task-irrelevant information. Accordingly, in A2T, cognitive control is conceptualized as the moment-to-moment competition (or differential influences) between the effects of contextual demands *vs.* the effects of representations in WM and the affect on internal selection likelihoods (see also Wegner, 1994; Wegner, Erber, & Zanakos, 1993). This means that successful control is a state in which the contextual demands component out-weighs/influences other factors, and vice-versa with respect to control failure (Awh et al., 2012). For example, following a highly arousing but task-irrelevant spontaneous memory (e.g., a car accident), the prepotent response is to continue analyzing information related to that memory (e.g., other memories from the event) (Klinger, 2013; McVay & Kane, 2013). However, the contextual demands signal may be strong enough to bias selection to task-relevant-, and memory event unrelated-, information driving a state of cognitive control (McVay & Kane, 2013).

Furthermore, A2T assumes that the influence of contextual demands on focused attention, relative to other components of the model, should be more stable and persist over longer time periods. Whereas thoughts (representations in WM) and their influence decay in seconds, due to constraints of short-term memory, contextual demands for sustained attention should persist for longer time periods. For example, the instructions and motivation to optimally perform on a demanding task is likely to remain the same for the duration of the task. The facilitation or interference of stimulus-driven signals (representations in WM and affect), rather than the momentary ability to exert focused attention (the sustained cognitive control/inhibition over distracting information), is what changes from moment-to-moment. Thus, momentary changes in cognitive control are **due to changes in stimulus-driven signals rather than fluctuations in goal-directed signaling** (McVay & Kane, 2012). This may seem trivial, but often in the cognitive control literature, if cognitive control fluctuations are not due to stimulus-driven fluctuations, then failures must be attributed to a sub-component of cognitive control—i.e., a "homunculus" control component (Baddeley, 2012; Everaert, Grahek, & Koster, 2017). The proposed A2T model enables examining the stability of cognitive control (or conversely failures) as the frequency of momentary states of attention directed toward/away from task-irrelevant information (i.e., task-unrelated thoughts)—as a function of the different states in context and individual differences of low and high reactivity. Importantly, A2T provides a conceptualization and computational operationalization of cognitive control without an extraneous/homunculus control mechanism (Hazy, Frank, & O'Reilly, 2007). Temporal fluctuations in (dys)control emerge from the complex temporal interactions of the low-level components of the A2T model, as well as the inherent stochasticity in internal attention selection and individual differences in feedback paths representing cognitive and affective reactivity.

Individual Differences in Cognitive (Dys)control. In addition to contextual demands, the momentary changes in selection bias likelihoods are also determined by affect, representations in WM, and reactivity paths, as well as the stochastic element of internal selection. Figure 9A illustrates how a simulated model with high reactivity has random elevations or spikes in the selection likelihood of bias to negative task-unrelated representations, even under high contextual demands for focused attention. Moreover, the model predicts that when the *task-relevant information is negatively valenced* and inevitably selected into WM, such information/representation in WM increases the probability of subsequent selections of *affect-congruent but potentially task-irrelevant* information (McVay & Kane, 2013). This likely leads to a greater frequency of cognitive control failures in the form of attending to task-unrelated thoughts. This prediction is in-line with several observed phenomenon, such as the Emotional Stroop effect, wherein the naming of emotional, *vs.* neutral, words are slowed (Williams, Mathews, & MacLeod, 1996). A2T provides an alternative, task-unrelated thoughts-based account, of the effect (cf. conflict-based account of Stroop effects) (Algom, Chajut, & Lev, 2004). Likewise, these A2T predictions are also in line with findings of increased task-unrelated thoughts and impaired external attention performance following (task-embedded) personal concern cues (McVay & Kane, 2013).

The ability to sustain attention on task-related information, in the presence of emotionally evocative task-relevant information, is crucial for functioning in various situations. For example, being able to focus in the presence of stressors, such as performing well during stressful interviews. A more extreme example is "first responders" who need to focus in the presence of potentially traumatic scenes. Figure 9B illustrates how high reactivity, together with negatively-valenced task-relevant information, increases the selection bias to

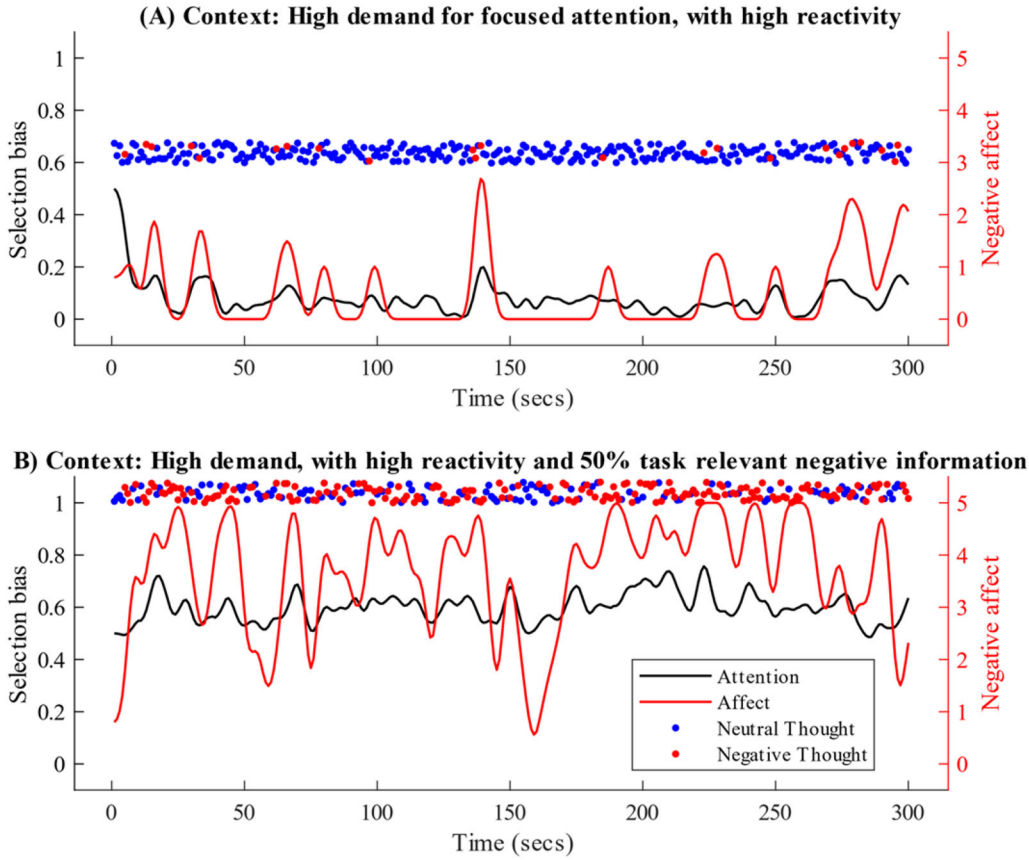

Figure 9. Simulation of trajectories under high-contextual demands (= 4) and high emotional and cognitive reactivity (=.25). In (A), all task-relevant information is neutrally valenced. In (B), half the information is negatively valenced (e.g., memorizing a list of mixed negative and neutral words).

negative representations (as the context requires) but has a stronger impact on variability and mean negative affect. As such, individual differences in the strength of critical paths—from representation in WM to affect and from affect to internal attentional selection—help to explain previously observed deficits in cognitive control over negatively valenced information in individuals vulnerable to depression and anxiety (Grahek et al., 2018; Hertel, 1997; LeMoult & Gotlib, 2018).

Summary: Cognitive (Dys)control in A2T

The A2T conceptualization of cognitive (dys)control and computational simulations offer several insights. For example, when studying the interactions of cognition and emotion on cognitive (dys)control, aggregating performance across time (e.g., across an entire task or task-block) may fail to capture the key signal/effects (Algom et al., 2004). Rather, it may be more conceptually accurate to make hypotheses and empirical tests on the momentary temporal level—e.g., to examine momentary affective changes as both antecedents (predictors) and consequences (outcome) of momentary control failures and vice-versa (Ruby et al., 2013; Ruimi et al., 2018; Smallwood & Schooler, 2015).

Interpretation Bias

Interpretation bias is the tendency for certain individuals to interpret ambiguous information as negative. Such biases are associated with depression and anxiety as well as other forms of cognitive biases (e.g., memory bias; Everaert et al., 2014) and vulnerabilities (e.g., repetitive negative thinking; Everaert et al., 2020; Hirsch et al., 2018). Accordingly, interpretation bias is a primary target for many psychological treatments for depression and anxiety as well as emerging cognitive training methodologies (Everaert, Podina, & Koster, 2017; Hirsch et al., 2018). Notably, a core feature of interpretation bias is the preferential selection of one interpretation over competing resolutions to ambiguous information (Gernsbacher & Faust, 1991).

Interpretation Bias in A2T

The process of interpretation bias is, by definition, triggered by information that is ambiguous or may be alternatively interpreted. Various interpretations of such an ambiguous situation and subsequent thoughts compete for selection (Everaert, 2021). In A2T, the specific interpretation selected (e.g., negative, positive, or neutral) is determined by the selection likelihoods (at *Time t*) influenced by the immediately preceding state of the system—such as the active representations in WM and affective state (at *Time* $t-1$). For example, imagine a person arriving at a social event and greeted by the host who has an ambiguous non-verbal expression (e.g., facial expression, body posture). If the person arrived with negative expectations (in the form of negative thoughts/content in WM) and/or in a negative mood, this then increases the likelihood that a negative

interpretation is selected for further processing (Davey, Bickerstaffe, & MacDonald, 2006), i.e., entering WM and reaching awareness. That is, people with *active* negative expectations/beliefs are more likely to demonstrate negative interpretation biases. This prediction is, however, in contrast to current perspectives on the roles of priming (whether conscious or unconcious, and conceptually similar to path *d* in Figure 1A) and affect as outcomes rather than causes of biased interpretations (MacLeod & Mathews, 2012; but see Halberstadt, Niedenthal, & Kushner, 1995 for evidence on the effects of affect on interpretation bias). To reconcile this disparity, we again highlight that the model emphasizes the transaction that unfolds from moment-to-moment in time (see Figure 1B) between attention, WM, and affect—i.e., they are reciprocal and circular causes- *and* outcomes- of one another in time. These A2T predictions are in line with recent findings (Everaert, 2021) and emerging theory regarding the interacting roles of multiple cognitive (attention, WM) biases in driving interpretation biases (Everaert et al., 2020).

A2T and Maladaptive Internally-Directed Cognition: Summary

A2T offers a single, unified explanatory framework for the emergence of various forms of (mal)adaptive IDC including common forms of maladaptive thinking. A2T delivers a computational account and basis to simulate the theorized dynamic system underlying IDC. Thus, A2T not only *describes* what internally-directed cognitive processes may become dysregulated but also provides a *computational* framework for how dysregulation emerges from what is also a normative/adaptive system. This enables quantifying trajectories of the A2T model under key differential conditions (i.e., contextual and individual differences factors) and comparing them to empirically observed phenomena of interest.

Future directions

This paper lays out our initial efforts to use the A2T model to better understand the role of internal attention in (mal)-adaptive IDC. Future work applying, testing, and further developing A2T could focus on several key questions and directions. First, A2T may help generate more refined hypotheses and quantitative predictions regarding the function of internal attention in IDC (Borsboom et al., 2021; Grahek et al., 2021; van Rooij & Baggio, 2021). A2T's computational framework enables simulating specific contexts and individual differences and their interactions. In turn, such simulations enable testing the overlap between simulated and empirically observed phenomena including behavior (van Vugt & van der Velde, 2018). Such computational approaches have been utilized in research fields, such as decision-making (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and external attention (Logan, 2004). Such work to-date has helped refine decision-making and visual attention theories through an iterative process of (mis)-matching theoretical predictions, simulated data, and observed behavior (Muthukrishna & Henrich, 2019). Using A2T, this type of computational approach may serve to iteratively refine theory and our understanding of internal attention and (mal)adaptive IDC in the coming years.

It is therefore important to reiterate that A2T is a *formal* model that adopts a dynamic systems approach; but is not a *data* model. A2T provides specific predictions/hypotheses about the behavior of IDC under different contexts and individual differences; but is not specifically designed for parametric fitting to experimental data. While this may be possible given well-matching data (e.g., high temporal resolution experiential momentary assessments) it was not the aim when developing A2T. Accordingly, support for A2T is currently limited to simulations demonstrating previously empirically observed effects and behaviors (e.g., high mind-wandering under low contextual demands), and A2T's significant value lies in its potential for providing a theoretically plausible explanation (van Rooij & Baggio, 2021) of (mal)adaptive IDC.

A second inter-related future direction for theory and A2T model development is the examination, refinement, and further elaboration of specific model components and pathways. For example, in this report and instantiation of A2T, our computational formulation assumes that WM storage capacity (of five items) is similar across individuals (models). However, there are meaningful differences between individuals in their WM capacity (Kane, Conway, Hambrick, & Engle, 2007). Furthermore, such differences in capacity have been theoretically and empirically linked to individual differences in key processes, such as attentional control and emotion regulation (Coifman et al., 2019; Kane et al., 2007). Future work may further specify individual differences in A2T's WM component to simulate and test their salutary and/or maladaptive effects on trajectories of the dynamic system (Welhaf et al., 2020). Furthermore, for conceptual and computational application and simplicity, we separated internal attentional and WM into the selection and short-term storage functions, respectively. However, internal attention and WM are highly inter-related processes (see *Working memory & internal attention* section). Future work may thus incorporate more elaborate rules for how these processes interact and accordingly produce more elaborate simulation events. For example, we adopt a "winner takes all" approach for the filtering/gatekeeping function of internal attention into WM. However, A2T may be modified for attenuator accounts (Treisman, 1964; Yantis & Johnston, 1990) of how internal attentional selection passes the information on to working-memory. Furthermore, although we focused here on *negative* affect in this first instantiation of A2T, *positive* affect and related working memory representations may also be readily integrated into the model. Likewise, A2T development may explore factors or intervention strategies that modulate inter-component pathways, including modulation of (affective) *reactivity to representations in WM* and (attentional) *reactivity to affective state*.

Third, A2T model components and pathways could also be adapted to be applied to a variety of additional phenomena characterized by IDC that unfold in time. For example,

A2T may be applied to model variability in moment-to-moment focus of attention during mindfulness meditation. Mindfulness meditation is a practice in which attention is effortfully directed to the person's internal experience in the present moment. Interestingly, in such a practice, the term "internal" is typically defined in a broader sense then we adopt here and includes attention to interoceptive and bodily sensations (Dixon et al., 2014; Lutz, Jha, Dunne, & Saron, 2015). For example, in focused attention meditation, a common practice entails "anchoring" attention to some sensory-perceptual object, such as one's breath (Lutz et al., 2015). To better understand how attention unfolds during such mindfulness practices, A2T may be used to model and simulate how internal attention drifts to thoughts and feelings (i.e., effects of affect and stochasticity of internal attentional selection) as well as volitional orienting attention back to one's breath sensations (i.e., effects of contextual demands for focused attention) over the course of mindfulness meditation (Hadash & Bernstein, 2019). To permit such applications, A2T may be generalized in several ways, such as adding interoceptive- and sensation-based information as competing objects for selective attention, or accounting for meta-cognitive monitoring processes, such as meta-awareness (Dunne, Thompson, & Schooler, 2019; Ruimi et al., 2018; Smallwood, 2013a).

Fourth, A2T may help us in thinking about how we conceptualize and formulate phenomena of interest, what important questions should be addressed, and accordingly what methodologies should be developed to answer these questions (Muthukrishna & Henrich, 2019). One key methodological challenge facing the study of IDC is the capacity for controlled experimental study and measurement of internal attentional processes (Amir et al., 2021; Posner, 2016). Whereas the methodological paradigm used to measure and quantify external attention allow for experimental control over the content, timing, and location of external (e.g., visual) stimuli, such control has been largely enigmatic for the study of internal events (e.g., thoughts) (Amir et al., 2021; Posner, 2016). WM tasks enable objective measurement of the processing of internal information but lack the capacity to sufficiently disentangle storage and attentional processes (Atkinson et al., 2018; Baddeley, 2012). For example, Amir et al. (2021) developed the Simulated Thoughts Paradigm to deliver idiographic stimuli that simulate the *content* and the *experience* of one's own verbal thoughts; and then embed these simulated thought stimuli in established attentional tasks to quantify internal attentional processes (e.g., internal selective attention bias, attentional disengagement from thoughts). We believe that similar methods designed to deliver experimental control over phenomenologically-valid stimuli that simulate internal experiences (e.g., imagery, autobiographical memories) may be critical to advance the study of internal attention broadly and the utility of A2T more specifically (see also Engen & Anderson, 2018; Nobre & Stokes, 2019).

A related methodological challenge facing the study of IDC relates to the temporal resolution of measurement methods used to quantify attention, working memory, and emotion. Indeed, typical current behavioral methodologies are not designed to quantify dynamic processes with a high temporal resolution, but rely on aggregated estimates of such processes by collapsing across repeated measures (e.g., trial-level) data (Zvielli, Bernstein, & Koster, 2015) or rely on less intensive measurement epochs (Smallwood & Schooler, 2015). Yet, A2T and simulations illustrate the importance of low-level moment-to-moment processes in understanding (mal)adaptive IDC. Accordingly, one future direction may focus on measuring and quantifying processes of interest, such as attention and working memory, as dynamic processes from moment-to-moment in time (e.g., trial-to-trial, continuous real-time) (Amir, Zvielli, & Bernstein, 2016; Amir et al., 2021; Schwartz & Stone, 2007). Such efforts may also be facilitated by the integration of continuous real-time multi-modal measurement technologies (e.g., eye-movement, pupil dilation, peripheral psychophysiology, electrophysiology, real-time functioning imaging technologies). A second approach to this challenge could explore down-scaling of A2T simulation data to quantify "meta"-parameters (e.g., aggregated mean selection bias, aggregated mean negative affect, the proportion of task-unrelated to task-related thoughts)—to better match the lower temporal scale of common current repeated-measures behavioral methods (Ruby et al., 2013). For example, A2T simulation data may be aggregated (e.g., mean, variability) over longer (e.g., 10-min) temporal epochs (e.g., see Figures 8A,B). This approach could, for example, enable estimates of the aggregated probability/frequency and duration of IDC processes, such as mindwandering episodes, as a function of context demands and individual differences in reactivity. Such meta-parameters could then be more directly contrasted with, for example, intensive experience sampling data (e.g., thought probes over a 10 min epoch) (Smallwood & Schooler, 2015; Welhaf et al., 2020) or cognitive-experimental behavioral task data (e.g., aggregated attention bias estimates over a 10 min epoch).

Summary

A2T is an explanatory formal model (Borsboom et al., 2021; Epstein, 2008; Grahek et al., 2021; van Rooij & Baggio, 2021). It provides an explanation—through a system of computational principles—of how higher-order phenomena of IDC (e.g., repetitive negative thinking) emerge. Descriptively, A2T converges with previous theories of IDC, such as the process-occurrence framework (Smallwood, 2013a) and dynamics of spontaneous thought (Christoff et al., 2016). One critical advancement of A2T is its computational formalization. This results in more explicit assumptions and mechanistic principles and thereby rigorous examination of the theory (van Rooij & Baggio, 2021). For example, the computational formalization of A2T permits contrasts of findings from model simulations relative to empirical observations and experimental findings related to IDC (e.g., reduced mindwandering under increased task demands). It thereby also facilitates direct and rigorous examination of the empirical verisimilitude of the theory

and model (Borsboom et al., 2021; Muthukrishna & Henrich, 2019; van Rooij & Baggio, 2021). Accordingly, we find that A2T helps account for various expressions of IDC (e.g., goal-directed thinking, repetitive thinking) as well as to test theory and hypotheses regarding the role of internal attention in (mal)adaptive IDC (Amir et al., 2021).

In summary, Attention-to-Thoughts (A2T) is a dynamic system and computational model of internal attention in (mal)adaptive IDC. A2T implicates internal attention at the functional center of a dynamic system of inter- and transacting cognitive and affective processes subserving IDC. Critically, A2T reflects the dynamic temporal nature of IDC through the conceptualization and computational formalization of the interactions of system components in time. Simulations over time of the model help to demonstrate how various forms of IDC (e.g., mindwandering, stream of thought) emerge from moment-to-moment interaction of A2T system components. A2T may also help to understand how various forms of maladaptive IDC (e.g., repetitive negative thinking) and thereby mental health problems emerge from the very same system that also subserves adaptive cognition. We hope that A2T may contribute to growing field-wide interest in internal attention as well as the potential and importance of research that seeks to conceptually and computationally reflect the dynamic complexity of mental phenomena that characterize human mental life.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

ORCID

Iftach Amir http://orcid.org/0000-0001-6005-8122

References

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931–965. doi:10.1037/bul0000021

Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional stroop phenomenon: A generic slowdown, not a stroop effect. *Journal of Experimental Psychology: General*, *133*(3), 323–338. doi:10.1037/0096-3445.133.3.323

Amir, I., Ruimi, L., & Bernstein, A. (2021). Simulating thoughts to measure and study internal attention in mental health. *Scientific Reports*, *11*(1), 2251. 10.1038/s41598-021-81756-w.

Amir, I., Zvielli, A., & Bernstein, A. (2016). (De)coupling of our eyes and our mind's eye: A dynamic process perspective on attentional bias. *Emotion. American Psychological Association*, *16*(7), 978–986. doi:10.1037/emo0000172

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Science*, *1316*, 29–52. doi:10.1111/nyas.12360

Atkinson, A. L., Berry, E. D. J., Waterman, A. H., Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2018). Are there multiple ways to direct attention in working memory? *Annals of the New York Academy of Sciences*, *1424*(1), 115–126. doi:10.1111/nyas.13634

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. doi:10.1016/j.tics.2012.06.010

Axelrod, V., Rees, G., & Bar, M. (2017). The default network and the combination of cognitive processes that mediate self-generated thought. *Nature Human Behaviour*, *1*(12), 896–910. doi:10.1038/s41562-017-0244-9

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29. doi:10.1146/annurev-psych-120710-100422

Bernstein, A., Hadash, Y., Lichtash, Y., Tanay, G., Shepherd, K., & Fresco, D. M. (2015). Decentering and related constructs: A critical review and metacognitive processes model. *Perspectives on Psychological Science*, *10*(5), 599–617. doi:10.1177/1745691615594577

Bor, D., & Seth, A. (2012). Consciousness and the Prefrontal Parietal Network: Insights from attention, working memory, and chunking. *Frontiers in Psychology*, *3*, 63. doi:10.3389/fpsyg.2012.00063

Borkovec, T. D., Robinson, E., Pruzinsky, T., & DePree, J. A. (1983). Preliminary exploration of worry: Some characteristics and processes. *Behaviour Research and Therapy*, *21*(1), 9–16. doi:10.1016/0005-7967(83)90121-3

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. doi:10.1177/1745691620969647

Buetti, S., & Lleras, A. (2016). Distractibility is a function of engagement, not task difficulty: Evidence from a new oculomotor capture paradigm. *Journal of Experimental Psychology: General*, *145*(10), 1382–1405. doi:10.1037/xge0000213

Buzsáki, G. (2006). *Rhythms of the brain*. New York, NY: Oxford University Press. doi:10.1093/acprof:Oso/9780195301069.001.0001

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: An attentional account. *Nature Reviews. Neuroscience*, *9*(8), 613–625. doi:10.1038/nrn2459

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8719–8724. doi:10.1073/pnas.0900234106

Christoff, K., Irving, Z. C., Fox, K. C. R. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews. Neuroscience*, *17*(11), 718–731. doi:10.1038/nrn.2016.113

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73–101. doi:10.1146/annurev.psych.093008.100427

Chun, M. M., & Johnson, M. K. (2011). Memory: Enduring traces of perceptual and reflective attention. *Neuron*, *72*(4), 520–535. doi:10.1016/j.neuron.2011.10.026

Coifman, K. G., Kane, M. J., Bishop, M., Matt, L. M., Nylocks, K. M., & Aurora, P. (2019). Predicting negative affect variability and spontaneous emotion regulation: Can working memory span tasks estimate emotion regulatory capacity? *Emotion*, *21*(2), 297–314. doi:10.1037/emo0000585

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Curci, A., Lanciano, T., Soleti, E., & Rimé, B. (2013). Negative emotional experiences arouse rumination and affect working memory capacity. *Emotion*, *13*(5), 867–880. doi:10.1037/a0032492

Davey, G. C. L., Bickerstaffe, S., & MacDonald, B. A. (2006). Experienced disgust causes a negative interpretation bias: A causal role for disgust in anxious psychopathology. *Behaviour Research and Therapy*, *44*(10), 1375–1384. doi:10.1016/j.brat.2005.10.006

Davidson, R. J. (1998). Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, *12*(3), 307–330. doi:10.1080/026999398379628

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. doi:10.1038/nature04766

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual. *Annual Review of Neuroscience*, *18*(1), 193–222. doi:10.1146/annurev.ne.18.030195.001205

Dixon, M. L., Fox, K. C. R., & Christoff, K. (2014). A framework for understanding the relationship between externally and internally directed cognition. *Neuropsychologia*, *62*, 321–330. doi:10.1016/j.neuropsychologia.2014.05.024

Dunne, J. D., Thompson, E., & Schooler, J. (2019). Mindful meta-awareness: Sustained and non-propositional. *Current Opinion in Psychology*, *28*, 307–311. doi:10.1016/j.copsyc.2019.07.003

Ehring, T., & Behar, E. (2020.) Transdiagnostic view on worrying and other negative mental content. In *Generalized anxiety disorder and worrying: A comprehensive handbook for clinicians and researchers*. (eds A.L. Gerlach and A.T. Gloster) Wiley-Blackwell. doi:10.1002/9781119189909.ch4

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International Journal of Cognitive Therapy*, *1*(3), 192–205. doi:10.1680/ijct.2008.1.3.192

Ehring, T., Zetsche, U., Weidacker, K., Wahl, K., Schönfeld, S., & Ehlers, A. (2011). The Perseverative Thinking Questionnaire (PTQ): Validation of a content-independent measure of repetitive negative thinking. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 225–232. doi:10.1016/j.jbtep.2010.12.003

Eich, E. (1995). Searching for mood dependent memory. *Psychological Science*, *6*(2), 67–75. doi:10.1111/j.1467-9280.1995.tb00309.x

Ellamil, M., Fox, K. C. R., Dixon, M. L., Pritchard, S., Todd, R. M., Thompson, E., & Christoff, K. (2016). Dynamics of neural recruitment surrounding the spontaneous arising of thoughts in experienced mindfulness practitioners. *NeuroImage*, *136*, 186–196. doi:10.1016/j.neuroimage.2016.04.034

Engen, H. G., & Anderson, M. C. (2018). Memory control: A fundamental mechanism of emotion regulation. *Trends in Cognitive Sciences*, *22*(11), 982–995. doi:10.1016/j.tics.2018.07.015

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12. Retrieved from http://jasss.soc.surrey.ac.uk/11/4/12.html

Epstein, R. (2000). The neural-cognitive basis of the Jamesian stream of thought. *Consciousness and Cognition*, *9*(4), 550–575. doi:10.1006/ccog.2000.0486

Everaert, J. (2021). Interpretation of ambiguity in depression. *Current Opinion in Psychology*, *41*, 9–14. doi:10.1016/j.copsyc.2021.01.003

Everaert, J., Bernstein, A., Joormann, J., & Koster, E. H. W. (2020). Mapping dynamic interactions among cognitive biases in depression. *Emotion Review*, *12*(2), 93–110. doi:10.1177/1754073919892069

Everaert, J., Duyck, W., & Koster, E. H. W. (2014). Attention, interpretation, and memory biases in subclinical depression: A proof-of-principle test of the combined cognitive biases hypothesis. *Emotion*, *14*(2), 331–340. doi:10.1037/a0035250

Everaert, J., Grahek, I., & Koster, E. H. (2017). Individual differences in cognitive control over emotional material modulate cognitive biases linked to depressive symptoms. *Cognition & Emotion*, *31*(4), 736–746. doi:10.1080/02699931.2016.1144562

Everaert, J., Podina, I. R., & Koster, E. H. W. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review*, *58*, 33–48. doi:10.1016/j.cpr.2017.09.005

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353. doi:10.1037/1528-3542.7.2.336

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. doi:10.1177/0963721410386677

Fox, K. C. R., Andrews-Hanna, J. R., Mills, C., Dixon, M. L., Markovic, J., Thompson, E., & Christoff, K. (2018). Affective neuroscience of self-generated thought. *Annals of the New York Academy of Sciences*, *1426*(1), 25–51. doi:10.1111/nyas.13740

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020. doi:10.1177/1745691617705892

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. doi:10.1016/j.tics.2011.11.014

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 245–262. doi:10.1037/0278-7393.17.2.245

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. doi:10.1016/j.jmp.2016.01.006

Grahek, I., Everaert, J., Krebs, R. M., & Koster, E. H. W. W. (2018). Cognitive control in depression: Toward clinical models informed by cognitive neuroscience. *Clinical Psychological Science*, *6*(4), 464–480. doi:10.1177/2167702618758969

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. *Perspectives on Psychological Science*, *16*(4), 803–815. doi:10.1177/1745691620966794

Hadash, Y., & Bernstein, A. (2019). Behavioral assessment of mindfulness: Defining features, organizing framework, and review of emerging methods. *Current Opinion in Psychology*, *28*, 229–237. doi:10.1016/j.copsyc.2019.01.008

Halberstadt, J. B., Niedenthal, P. M., & Kushner, J. (1995). Resolution of lexical ambiguity by emotional state. *Psychological Science*, *6*(5), 278–282. doi:10.1111/j.1467-9280.1995.tb00511.x

Harvey, A. G., Watkins, E., & Mansell, W. (2004). *Cognitive behavioural processes across psychological disorders: A transdiagnostic approach to research and treatment*. New York, NY: Oxford University Press.

Haslbeck, J., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019). Modeling psychopathology: From data models to formal theories.

Hayes, S., Hirsch, C., & Mathews, A. (2008). Restriction of working memory capacity during worry. *Journal of Abnormal Psychology*, *117*(3), 712–717. doi:10.1037/a0012908

Hazy, T. E., Frank, M. J., & O'reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613. doi:10.1098/rstb.2007.2055

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling BT. In B. U. Forstmann & E.-J. Wagenmakers (Eds.) *An introduction to model-based cognitive neuroscience* (pp. 25–48). New York, NY: Springer New York. doi:10.1007/978-1-4939-2236-9_2

Hertel, P. T. (1997). On the contributions of deficent cognitive control to memory impairments in depression. *Cognition & Emotion*, *11*(5–6), 569–583. doi:10.1080/026999397379890a

Hirsch, C. R., Krahé, C., Whyte, J., Loizou, S., Bridge, L., Norton, S., & Mathews, A. (2018). Interpretation training to target repetitive negative thinking in generalized anxiety disorder and depression. *Journal of Consulting and Clinical Psychology*, *86*(12), 1017–1030. doi:10.1037/ccp0000310

Hitchcock, C., Golden, A. M. J., Werner-Seidler, A., Kuyken, W., & Dalgleish, T. (2018). The impact of affective context on autobiographical recollection in depression. *Clinical Psychological Science*, *6*(3), 315–324. doi:10.1177/2167702617740672

Hollingworth, A., & Luck, S. J. (2009). The role of visual working memory (VWM) in the control of gaze during visual search. *Attention, Perception & Psychophysics*, *71*(4), 936–949. doi:10.3758/APP.71.4.936

Hong, R. Y., & Cheung, M. W.-L. (2015). The structure of cognitive vulnerabilities to depression and anxiety: Evidence for a common core etiologic process based on a meta-analytic review. *Clinical Psychological Science*, *3*(6), 892–912. doi:10.1177/2167702614553789

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. doi:10.1038/nn.4238

Ingram, R. E. (1990). Self-focused attention in clinical disorders. *Psychological Bulletin*, *107*(2), 156–176. doi:10.1037/0033-2909.107.2.156

James, W. (1890). The principles of psychology (Vol. II). New York, NY: Henry Holt and Company. doi:10.1037/11059-000

Joormann, J., & Gotlib, I. H. (2008). Updating the contents of working memory in depression: Interference from irrelevant negative material. *Journal of Abnormal Psychology*, *117*(1), 182–192. doi:10.1037/0021-843X.117.1.182

Juarrero, A. (2000). Dynamics in action: Intentional behavior as a complex system. *Emergence*, *2*(2), 24–57. doi:10.1207/S15327000EM0202_03

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In *Variation in working memory*. New York, NY: Oxford University Press.

Kanske, P., Plitschka, J., & Kotz, S. A. (2011). Attentional orienting towards emotion: P2 and N400 ERP effects. *Neuropsychologia*, *49*(11), 3121–3129. doi:10.1016/j.neuropsychologia.2011.07.022

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932. doi:10.1126/science.1192439

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–242. doi:10.3758/s13423-012-0359-y

Klinger, E. (1978). Modes of normal conscious flow. In K. S. Pope & J. L. Singer (Eds.), *The stream of consciousness: Scientific investigations into the flow of human experience* (pp. 225–258). Boston, MA: Springer US. doi:10.1007/978-1-4684-2466-9_9

Klinger, E. (2013). Goal Commitments and the content of thoughts and dreams: Basic principles. *Frontiers in Psychology*, *4*, 415. doi:10.3389/fpsyg.2013.00415

Klinger, E., Marchetti, I., & Koster, E. H. W. (2018). Spontaneous thought and goal pursuit: From functions such as planning to dysfunctions such as rumination. In *The Oxford handbook of spontaneous thought: Mind-wandering, creativity, and dreaming* (pp. 215–247). Oxford, UK: Oxford University Press.

Koster, E. H. W., De Lissnyder, E., Derakshan, N., & De Raedt, R. (2011). Understanding depressive rumination from a cognitive science perspective: The impaired disengagement hypothesis. *Clinical Psychology Review*, *31*(1), 138–145. doi:10.1016/j.cpr.2010.08.005

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468. doi:10.1037/0096-1523.21.3.451

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(10), E2016–E2025. doi:10.1073/pnas.1619316114

LeMoult, J., & Gotlib, I. H. (2019). Depression: A cognitive perspective. *Clinical psychology Review*, *69*, 51–66. doi:10.1016/j.cpr.2018.06.008

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *The Behavioral and Brain Sciences*, *28*(2), 169–194. doi:10.1017/S0140525X0500004X

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology*, *55*(1), 207–234. doi:10.1146/annurev.psych.55.090902.141415

Lutz, A., Jha, A. P., Dunne, J. D., & Saron, C. D. (2015). Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *American Psychologist*, *70*(7), 632–658. doi:10.1037/a0039585

MacLeod, C., & Mathews, A. (2012). Cognitive bias modification approaches to anxiety. *Annual Review of Clinical Psychology*, *8*, 189–217. doi:10.1146/annurev-clinpsy-032511-143052

Mansell, W., & McEvoy, P. M. (2017). A test of the core process account of psychopathology in a heterogenous clinical sample of anxiety and depression: A case of the blind men and the elephant? *Journal of Anxiety Disorders*, *46*, 4–10. doi:10.1016/j.janxdis.2016.06.008

Marchetti, I., Loeys, T., Alloy, L. B., & Koster, E. H. W. (2016). Unveiling the structure of cognitive vulnerability for depression: Specificity and overlap. *PLOS One*, *11*(12), e0168612. doi:10.1371/journal.pone.0168612

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin*, *136*(2), 188–197. doi:10.1037/a0018298

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "D'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 525–549. doi:10.1037/a0025896

McVay, J. C., & Kane, M. J. (2013). Dispatching the wandering mind? Toward a laboratory method for cuing "spontaneous" off-task thought. *Frontiers in Psychology*, *4*, 570. doi:10.3389/fpsyg.2013.00570

Mrkva, K., Westfall, J., & Van Boven, L. (2019). Attention drives emotion: Voluntary visual attention increases perceived emotional intensity. *Psychological Science*, *30*(6), 942–954. doi:10.1177/0956797619844231

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. doi:10.1038/s41562-018-0522-1

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449–461. doi:10.1016/j.tics.2017.03.010

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting attention to locations in perceptual versus mental representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–373. doi:10.1162/089892904322926700

Nobre, A. C., & Stokes, M. G. (2019). Premembering experience: A hierarchy of time-scales for proactive attention. *Neuron*, *104*(1), 132–146. doi:10.1016/j.neuron.2019.08.030

Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. *Journal of Personality and Social Psychology*, *61*(1), 115–121. doi:10.1037/0022-3514.61.1.115

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400–424. doi:10.1111/j.1745-6924.2008.00088.x

Oberauer, K. (2009). Design for a working memory. In *The psychology of learning and motivation* (Vol. 51, pp. 45–100). San Diego, CA: Academic Press. doi:10.1016/S0079-7421(09)51002-X

Oberauer, K., & Hein, L. (2012). Attention to Information in Working Memory. *Current Directions in Psychological Science*, *21*(3), 164–169. doi:10.1177/0963721412444727

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466–478. doi:10.1037/0096-3445.130.3.466

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews. Neuroscience*, *9*(2), 148–158. doi:10.1038/nrn2317

Poerio, G. L., Totterdell, P., & Miles, E. (2013). Mind-wandering and negative mood: Does one thing really lead to another? *Consciousness and Cognition*, *22*(4), 1412–1421. doi:10.1016/j.concog.2013.09.012

Posner, M. (2016). Orienting of attention: Then and now. *Quarterly Journal of Experimental Psychology*, *69*(10), 1864–1875. doi:10.1080/17470218.2014.937446

Posner, M. I. (1994). Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(16), 7398–7403. doi:10.1073/pnas.91.16.7398

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 676–682. doi:10.1073/pnas.98.2.676

Ruby, F. J. M., Smallwood, J., Engen, H., & Singer, T. (2013). How self-generated thought shapes mood—The relation between mind-wandering and mood depends on the socio-temporal content of thoughts. *PLOS One*, *8*(10), e77554. doi:10.1371/journal.pone.0077554

Ruimi, L., Hadash, Y., Zvielli, A., Amir, I., Goldstein, P., & Bernstein, A. (2018). Meta-awareness of dysregulated emotional attention. *Clinical Psychological Science*, *6*(5), 658–670. doi:10.1177/2167702618776948

Scher, C. D., Ingram, R. E., & Segal, Z. V. (2005). Cognitive reactivity and vulnerability: Empirical evaluation of construct activation and cognitive diatheses in unipolar depression. *Clinical Psychology Review*, *25*(4), 487–510. doi:10.1016/j.cpr.2005.01.005

Schwartz, J. E., & Stone, A. A. (2007). The analysis of real-time momentary data: A practical guide. In *The science of real-time data capture: Self-reports in health research* (pp. 76–113). Oxford, UK: Oxford University Press.

Schweizer, S., & Dalgleish, T. (2016). The impact of affective contexts on working memory capacity in healthy populations and in individuals with PTSD. *Emotion*, *16*(1), 16–23. doi:10.1037/emo0000072

Seli, P., Konishi, M., Risko, E. F., & Smilek, D. (2018). The role of task difficulty in theoretical accounts of mind wandering. *Consciousness and Cognition*, *65*, 255–262. doi:10.1016/j.concog.2018.08.005

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology*, *4*(4), 569–579. doi:10.1016/0959-4388(94)90059-0

Sheppes, G., Suri, G., & Gross, J. J. (2015). Emotion regulation and psychopathology. *Annual Review of Clinical Psychology*, *11*(1), 379–405. doi:10.1146/annurev-clinpsy-032814-112739

Siemer, M. (2005). Mood-congruent cognitions constitute mood experience. *Emotion*, *5*(3), 296–308. doi:10.1037/1528-3542.5.3.296

Smallwood, J. (2013a). Distinguishing how from why the mind wanders: A process–occurrence framework for self-generated mental activity. *Psychological Bulletin*, *139*(3), 519–535. doi:10.1037/a0030010

Smallwood, J. (2013b). Searching for the elements of thought: Reply to Franklin, Mrazek, Broadway, and Schooler (2013). *Psychological Bulletin*, *139*(3), 542–547. doi:10.1037/a0031019

Smallwood, J., & Andrews-Hanna, J. (2013). Not all minds that wander are lost: The importance of a balanced perspective on the mind-wandering state. *Frontiers in Psychology*, *4*, 441. doi:10.3389/fpsyg.2013.00441

Smallwood, J., Fitzgerald, A., Miles, L. K., & Phillips, L. H. (2009). Shifting moods, wandering minds: Negative moods lead the mind to wander. *Emotion*, *9*(2), 271–276. doi:10.1037/a0014855

Smallwood, J., & Schooler, J. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *SSRN*, *66*, 487–518. doi:10.1146/annurev-psych-010814-015331

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, *132*(6), 946–958. doi:10.1037/0033-2909.132.6.946

Smith, P. L., & Sewell, D. K. (2013). A competitive interaction theory of attentional selection and decision making in brief, multielement displays. *Psychological Review*, *120*(3), 589–627. doi:10.1037/a0033140

Teasdale, J. D., & Barnard, P. J. (1993). Affect, cognition, and change: Re-modelling depressive thought. In *Affect, cognition, and change: Re-modelling depressive thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Teasdale, J. D., Dritschel, B. H., Taylor, M. J., Proctor, L., Lloyd, C. A., Nimmo-Smith, I., & Baddeley, A. D. (1995). Stimulus-independent thought depends on central executive resources. *Memory & Cognition*, *23*(5), 551–559. doi:10.3758/BF03197257

Theeuwes, J. (2019). Goal-driven, stimulus-driven, and history-driven selection. *Current Opinion in Psychology*, *29*, 97–101. doi:10.1016/j.copsyc.2018.12.024

Thurner, S., Hanel, R., & Klimek, P. (2018). *Introduction to the theory of complex systems*. New York, NY: Oxford University Press.

Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, *3*(6), 449–459. doi:10.1016/S0022-5371(64)80015-3

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nature Reviews. Neuroscience*, *16*(1), 55–61. doi:10.1038/nrn3857

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. doi:10.1177/1745691620970604

van Schie, K., & Anderson, M. C. (2017). Successfully controlling intrusive memories is harder when control must be sustained. *Memory*, *25*(9), 1201–1216. doi:10.1080/09658211.2017.1282518

van Vugt, M., & van der Velde, M. (2018). How does rumination impact cognition? A first mechanistic model. *Topics in Cognitive Science*, *10*(1), 175–191. doi:10.1111/tops.12318

Verschooren, S., Schindler, S., De Raedt, R., & Pourtois, G. (2019). Switching attention from internal to external information processing: A review of the literature and empirical support of the resource sharing account. *Psychonomic Bulletin & Review*, *26*(2), 468–490. doi:10.3758/s13423-019-01568-y

Vuilleumier, P., Armony, J., & Dolan, R. (2003). Reciprocal links between emotion and attention. *Human Brain Function*, *2*, 419–444.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, *134*(2), 163–206. doi:10.1037/0033-2909.134.2.163

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological review*, *101*(1), 34–52. doi:10.1037/0033-295x.101.1.34

Wegner, D. M., Erber, R., & Zanakos, S. (1993). Ironic processes in the mental control of mood and mood-related thought. *Journal of Personality and Social Psychology*, *65*(6), 1093–1104. doi:10.1037//0022-3514.65.6.1093

Welhaf, M. S., Smeekens, B. A., Gazzia, N. C., Perkins, J. B., Silvia, P. J., Meier, M. E., … Kane, M. J. (2020). An exploratory analysis of individual differences in mind wandering content and consistency. *Psychology of Consciousness: Theory, Research, and Practice*, *7*(2), 103–125. doi:10.1037/cns0000180

Whitmer, A. J., & Gotlib, I. H. (2013). An attentional scope model of rumination. *Psychological Bulletin*, *139*(5), 1036–1061. doi:10.1037/a0030923

Williams, J. M. G., Barnhofer, T., Crane, C., Herman, D., Raes, F., Watkins, E., & Dalgleish, T. (2007). Autobiographical memory specificity and emotional disorder. *Psychological Bulletin*, *133*(1), 122–148. doi:10.1037/0033-2909.133.1.122

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional stroop task and psychopathology. *Psychological Bulletin*, *120*(1), 3–24. doi:10.1037/0033-2909.120.1.3

Yantis, S., & Johnston, J. C. (1990). On the locus of visual selection: Evidence from focused attention tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 135–149. doi:10.1037/0096-1523.16.1.135

Zetsche, U., Bürkner, P. C., & Schulze, L. (2018). Shedding light on the association between repetitive negative thinking and deficits in cognitive control–a meta-analysis. *Clinical psychology Review*, *63*, 56–65. doi:10.1016/j.cpr.2018.06.001

Ziegler, D. A., Janowich, J. R., & Gazzaley, A. (2018). Differential impact of interference on internally- and externally-directed attention. *Scientific Reports*, *8*(1), 1–10. doi:10.1038/s41598-018-20498-8

Zvielli, A., Bernstein, A., & Koster, E. H. W. (2015). Temporal dynamics of attentional bias. *Clinical Psychological Science*, *3*(5), 772–788. doi:10.1177/2167702614551572