A Conceptual and Computational Dynamic Systems Theory of Internal Attention, Thought, and Emotion, & Mental Health: Attention-to-Thought (A2T) Model

Iftach Amir, M.A. & Amit Bernstein, Ph.D.

University of Haifa, Israel

Author note:

# Abstract

From moment-to-moment, our minds are challenged to process events in the external environment as well as internal events within our mind including our thoughts, memories and emotions. Notably, the propensity to focus inward is fundamental to the subjective experience of human mental life, adaptive functioning and wellbeing, and accordingly also suffering and mental health. To understand how internal attention is linked to mental health we need to understand how it functions as part of a dynamic system of cognitive processes that sub-serve (mal)adaptive mental habits and related cognitive vulnerability. We therefore introduce a mechanistic model of the nature and functions of internal attention in the context or environmental demands of cognitions of past and future events and how individual differences within the model account for individual differences in higher-level cognitive vulnerabilities. The model characterizes how internal attention is expressed and iteratively transacts with contextual demands, representations in working-memory and affect, from moment-to-moment in time and context, and thereby sub-serves common forms of cognitive vulnerability as well as XXXXX and YYYY. We (1) first provide brief overview and definition of key concepts, including external attention, internal attention and working memory, and their functional relations to mental health; (2) outline the model and the common temporal trajectories in which the model may unfold; (3) how such trajectories may underly higher-order processes of cognitive vulnerabilities (e.g., repetitive negative thinking) or resilience (e.g., mindfulness) to mental ill health; (4) the utilities of the proposed model, such as for formulating computational models of XXXX; as well as (5) limitations of the proposed model and future directions for study of XXXX.

# Introduction

*“And in the naked light I saw,*

*Ten thousand people maybe more,*

***People talking without speaking,  
People hearing without listening,****”*

*Sound of Silence*, Simon & Garfunkel

Our propensity to direct our attention inwards – onto our thoughts, memories and feelings – is quintessential to human mental life and conscious experience. In addition to directing our attention externally out onto the world around us, we spend as much as ~30-50% with our attention turned inwards, “looking in” (Killingsworth & Gilbert, 2010; Klinger, 1978). This phenomenon is interesting and functionally important in that, among other functions, internal attention determines, and may be determined by, what internal information reaches awareness, including the selection of thoughts, memories, images and related internal experiences. Whereas some of our looking in can enable adaptive thinking (e.g., goal-setting, problem solving, mental time travel), it can also potentiate various forms of maladaptive thinking (e.g., spirals of negative repetitive thinking). Yet, our understanding of *how*, *when* and *for whom* higher-level (mal)adaptive thinking emerge from internal attentional selection, is limited and fragmented. Accordingly, we introduce the *Attention-to-Thoughts* (A2T) model – a conceptual and computational dynamic systems model of internal attention. 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 the emergence of (mal)adaptive thinking.

## External & 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 processing (*modulation*) (Chun, Golomb, & Turk-Browne, 2010; Desimone & Duncan, 1995).

*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). External attentional selection and modulation are governed by stimulus-driven (bottom-up) and goal-directed (top-down) systems (Chun et al., 2010; Corbetta & Shulman, 2002; Desimone & Duncan, 1995). The goal-directed system maintains active task sets and accordingly maps sensory representations with the relevant behavioral responses. This goal-directed- or attentional control- system is active in novel, relatively non-practiced and typically effortful behaviors (Shiffrin & Schneider, 1977). In contrast, the bottom-up stimulus-driven system is more automatic and reflexive, mediating the ability to reorient quickly to salient (e.g., novel, rewarding, threatening) events. It thus acts as an “alerting mechanism” or “circuit breaker” – disrupting top-down goal-directed processing whenever salient objects appear outside the focus of attention (Corbetta & Shulman, 2002; Moore & Zirnsak, 2017). Notably, some recent scholarship has critiqued this functional systems dichotomy arguing XXXXX (CITE).

*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., 2010; 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., 2010; 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., 2010). 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) and contextual constraints on thinking (Christoff et al., 2016).

**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. XX; Chun et al., 2010) |
| *External attention* | “Selection and modulation of sensory information” (p. XX, Chun et al., 2010) |
| *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 |

## Working memory & internal attention

Although there has been limited study of *internal* attention per se (Chun et al., 2010; 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, 2011; 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, XXX, YYYY) subserving WM (Baddeley, 2011; Oberauer, 2009), .

## Internal Attention, Working Memory and Awareness

One critical function of internal attention and working-memory is in subserving the “mental space” in-which thinking occurs – awareness. Attentional selection drives the gating or filtering of information that reaches conscious experience (Posner, 1994) and accordingly determines whether the focus of our awareness is directed externally or inwardly (Verschooren, Schindler, De Raedt, & Pourtois, 2019), as well as which of the vast number of external or internal sources of information enters conscious awareness at any given moment (CITE; van Vugt et al., 2018). Thus, just as external attention determines which information from the environment reaches awareness, internal attention determines which internal representations reach awareness. Essentially, attention and WM serve as key mechanisms subserving stream-of-thought. Attention gates, from moment-to-moment, which information reaches WM; and WM serves as the temporary storage mechanism of conscious detail (Bor & Seth, 2012).

## A dynamic and complex system

## Phenomenologically, thinking is not typically experienced as a series of discrete momentary states. Thoughts and memories unfold over time, they trigger other related thoughts, memories and emotions, and so forth. THEN IN THIS PARAGRAPH YOU INTRODUCE STABILITY/VARIABILITY.

## Moreover, COMPLEX. YOU INTRODUCE INTER-TRANSACTING FACTORS INCLUDING EMOTION.

## Temporal dynamics of internal attention

Phenomenologically, thinking is not typically experienced as a series of discrete momentary states. Thoughts and memories unfold over time, they trigger other related thoughts, memories and emotions, and so forth. This phenomenological stream-of-thought is typically experienced coherent in terms of the stability of its content or meaning (i.e., XXXXXX) yet it also transitive (i.e., XXXXXX) (Epstein, 2000; Welhaf et al., 2019). There are necessary shifts or switches in thought content – dictated by internal or external demands (Axelrod, Rees, & Bar, 2017; Christoff et al., 2016) as well as seemingly random shifts in thought content (e.g., XXXX) (Cabeza et al., 2008; Ellamil et al., 2016; Engen & Anderson, 2018).

The systemic complexity of context, cognition and emotion in which internal attentional processes operate is a major barrier to meaningfully answering these questions and thereby understanding temporal dynamics of internal attention or what guides internal attentional selection from moment-to-moment in time. Indeed, the processes affecting attention are also themselves influenced by attention - they have a circular causality (Kelso, 1995; Lewis, 2005). For example, attention gates the information entering WM, yet the contents or activated representations in WM also guide 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), yet emotional states amplify attentional selection by facilitating processing of- or disrupting on-going task and reorienting to- emotionally relevant information (Öhman, Flykt, & Esteves, 2001; Smallwood, Fitzgerald, Miles, & Phillips, 2009; Vuilleumier, Armony, & Dolan, 2003).

## A dynamic systems framework to model XXXX

Accordinglyu, we propose that a dynamic systems framework may be useful to advance our understanding of the nature and function of internal attention, and its interactions with other cognitive and emotion processes. 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 develops and unfolds over time), and have non-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” stream-of-thought emerge (Smallwood, 2013) from “lower-order” processes – e.g., the influence of selection of internal experiences into WM and awareness influence emotion, the influence of the contents of awareness and current emotional state on attentional selection? By adapting a dynamic systems approach, we therefore may be able to, building on understanding of how the basic low-level processes or components of the system dynamically inter- and trans- act, and understand how higher-level processes emerge (e.g., mindwandering, cognitive vulnerabilities). Accordingly, in recent years, this dynamic system approach has been similarly applied to theories of emotion, psychopathology (e.g., as symptom networks) and change processes during psychotherapy with various important insights yielded (Fried & Cramer, 2017; Hayes & Strauss, 1998; Hofmann, Curtiss, & Hayes, 2020; Lewis, 2005).

## Dynamic systems model of internal attention thinking and feeling

Thus, we propose that we should work to better understand how internal attention inter- and trans-acts with systematically proximal cognitive and emotion processes as a dynamic system that unfolds from moment to moment in time. Despite the theorized centrality of internal attention for thinking and feeling, we do not have such a dynamic model. Accordingly, we have limited capacity to model and predict (1) what processes determine the current momentary focus of internal attention (i.e., what drives selection of specific thought content into WM at any given moment in time?); and (2) the temporal stability and variability in thought content from moment-to-moment over time, and thus variability in stream-of-thought.

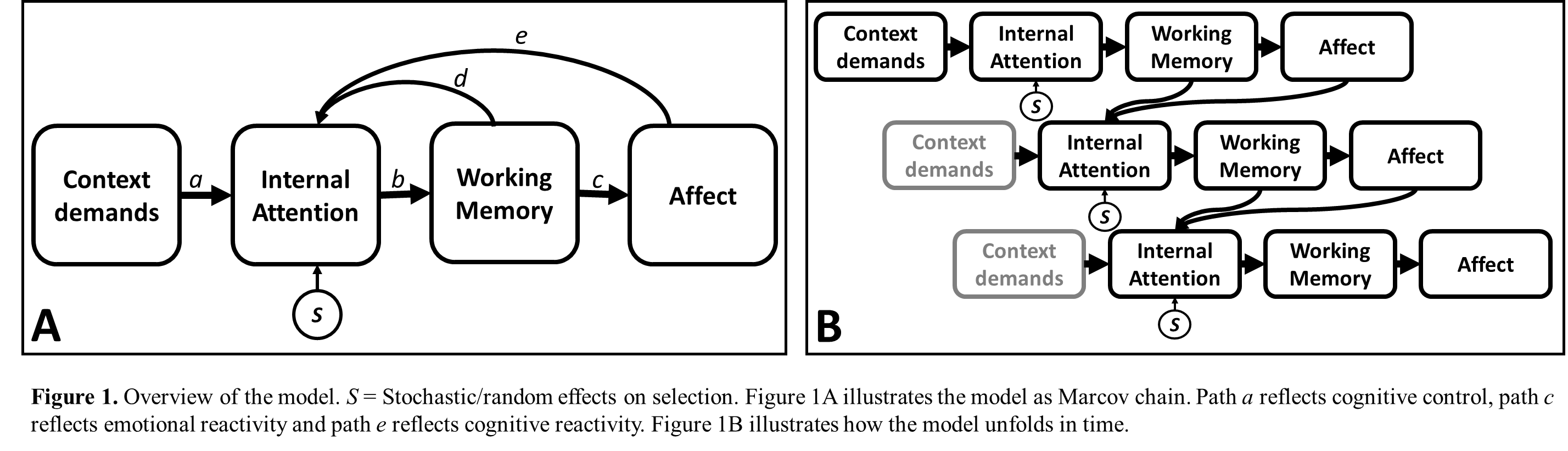
Moreover, in addition to these basic science insights about internal attention in cognition and emotion, such a model may have important applications for guiding study and advancing understanding of complex higher-level phenomena for which internal attention may be functionally central – such as cognitive vulnerability for mental health problems (for e.g., rumination and repetitive negative thinking), as well as phenomena such as mindwandering, mindfulness and spontaneous thought.

# ­Attention-to-Thoughts Model

*“All entities move and nothing remains still.”* – Heraclitus.

The A2T is a conceptual and computational dynamic systems model of internal attention and thought. A2T characterizes how internal attention affects and iteratively transacts with working-memory and affect processes, from moment-to-moment in time, as a function of contextual demands on attention. The model focuses on two key 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 determine for each moment in time: (i) the current internal attention selection likelihood or “bias” to attend to specific representations (e.g., negative vs. neutral representations), (ii) the current representations active in WM, and (iii) the current affective state (e.g., degree of momentary negative affect). Critically, each *momentary state* (Time *n*) is influenced by the previous (Time *n-1*), and affects the subsequent (Time *n+1*), momentary state. As such, 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). As such, the A2T aims to (a) conceptually and computationally define *momentary states* of this dynamic system of attention, thought and emotion, and (b) simulate and predict differential *trajectories* of this dynamic system.

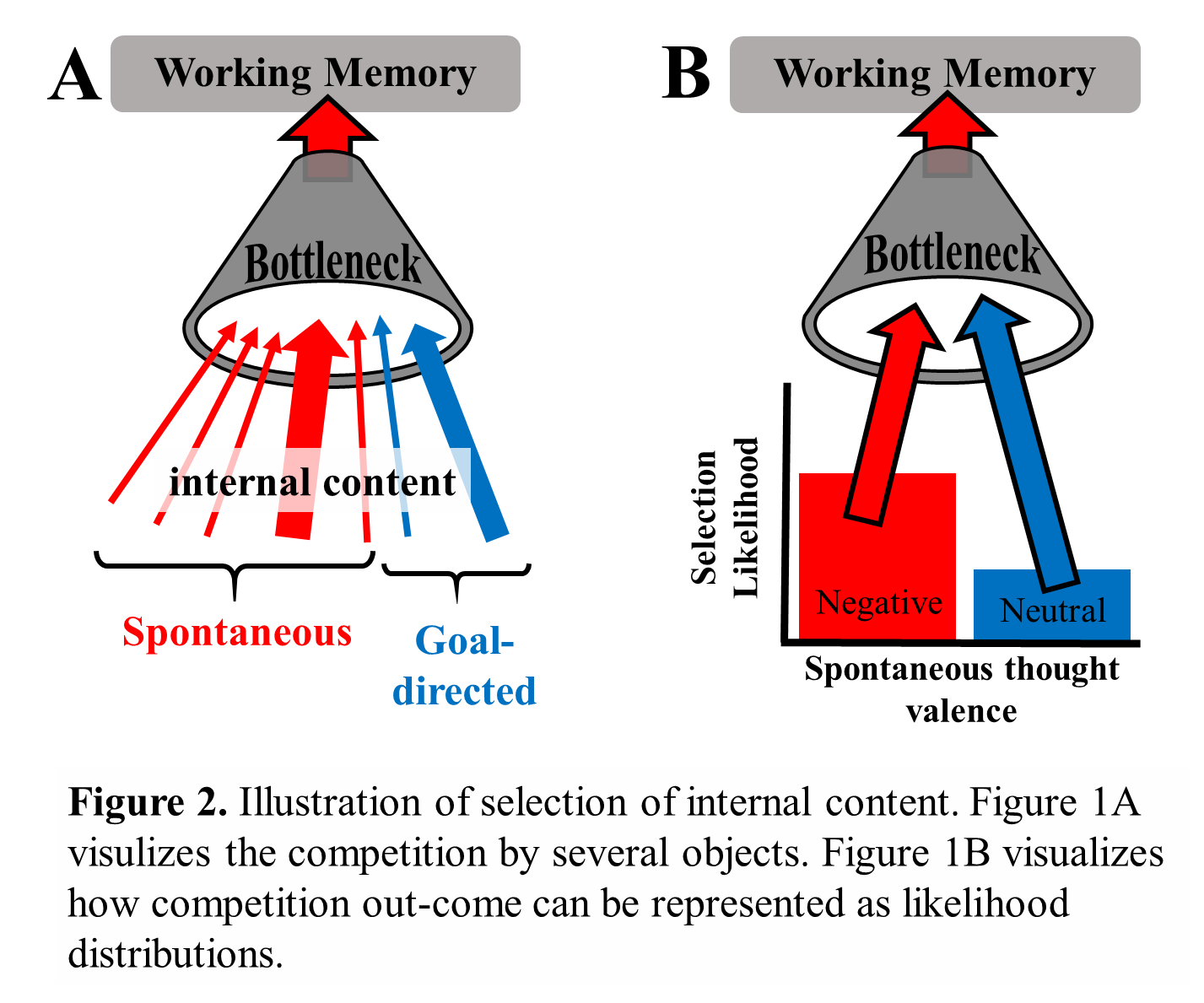
Utilizing A2T, we sought to characterize the following: (a) how internal attention selection is expressed from moment-to-moment; (b) how internal attention drives and iteratively transacts with WM, affective states, and how such components of the dynamic system feedback onto and thereby influence internal attentional selection of thought content; (c) how high and low contextual demands for focused attention to task-relevant information may moderate relations of internal attention, WM and affect; and (d) how individual differences in magnitude of influence between WM, affect and internal attention affect trajectories of the system. ingee and XXXX. To illustrate, we focus on higher-level cognition central to mental health. Specifically, using A2T, we illustrate may forms vulnerability; and b) how does A2T model account for these phenomena?



|  |  |  |  |
| --- | --- | --- | --- |
| 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) |
| 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 |
| Episodes of potential risk | Emergent characteristic | Hours, days | Trajectories |
| Vulnerability | Emergent characteristic | Days, months, years | Chronicity of trajectory or states in time, over longer time epics and across contexts |

## Model synopsis

First, the proposed A2T model characterizes the dynamic, inter- and trans-action, from moment-to-moment in time, of four key cognitive-affective components. These components constitute the repeated momentary process of selectively attending to internal representations into working-memory, such as verbal thoughts and memories (Alderson-Day & Fernyhough, 2015; Andrews-hanna, Smallwood, & Spreng, 2014; Dixon et al., 2014; Kanske, Plitschka, & Kotz, 2011), and their influence on affective state (LeDoux & Brown, 2017). The model components[[1]](#footnote-1) 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).

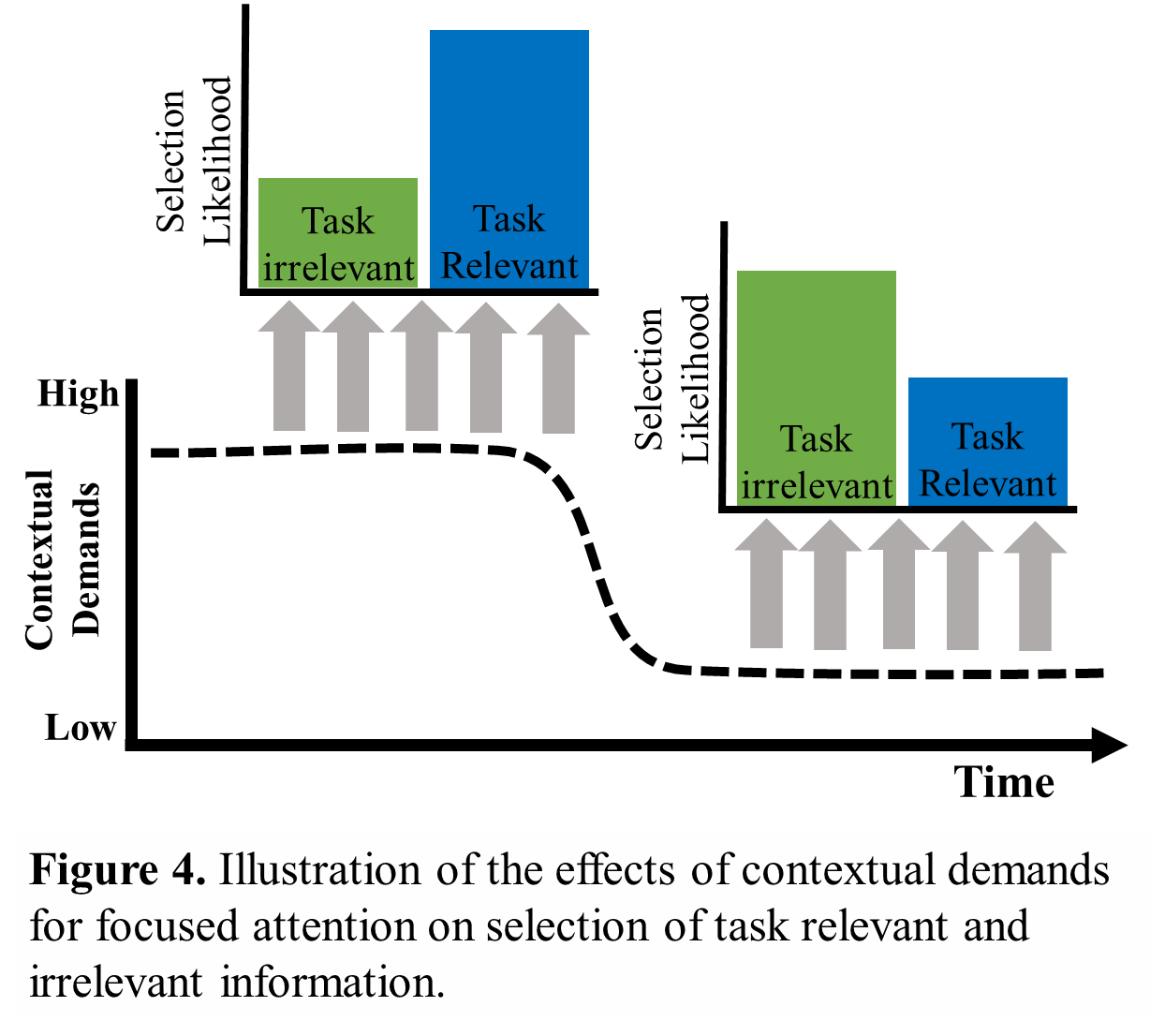


## Model components

**First, *internal attentional selection*** is at the functional center of the proposed dynamic system of attention, thought and emotion and thus the A2T model.At each moment in time, various sources for internal information, such as stimulus-driven spontaneous associative thoughts and memories as well as goal-directed thoughts, compete for processing. The selection determines what internal information receives 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., 2013)). 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]](#footnote-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 momentary states of the other components (their mechanism of influence is explained below). An additional source of influence on selection are stochastic/random factors such as random neural (system) noise (Buzsáki, 2006; Shadlen & Newsome, 1994; van Vugt et al., 2018). This randonmness 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 suchas 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, 2011)[[3]](#footnote-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), similar to the biasing effects of selection history on external attention (Awh, Belopolsky, & Theeuwes, 2012). See Figure 1A. This momentary selection bias is driven by a mechanism of spreading activation along a network of associations related to the representations active in working memory (Anderson & Lebiere, 1998; Oberauer, 2009).

**Third,** the **affect** component reflects the consequence(s) of the representation in working memory on current negative affective state (i.e., the more negative representations in WM trigger more negative affect; (Engen & Anderson, 2018; Ruby et al., 2013; Siemer, 2005). Like the ***representations in working memory*** component of the model, affect has phasic influence over likelihoods of selection by biasing selection in favor of affect-congruent internal representations. This similarly occurs 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)



**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, 2013) (see Figure 4). For example, external contextual high-demands for focused attention may be participating in a cognitively demanding 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 mind-wandering states (Christoff et al., 2016). Within the model, greater demand is translated into greater constraints the momentary attentional selection bias to negative representations (i.e., more contextual demands reduces internal selection bias driven by the current emotional state of the system) (Christoff et al., 2016; Whitmer & Gotlib, 2013). Similar to the influence of representations in WM, contextual demands bias selection by the spreading activation of associative networks related to the specific features of (the current) task-related representations (Luck & Vogel, 1997; Oberauer, 2009; Wolfe & Utochkin, 2019). For example, if the goal is to remember events from this morning, activation spreads to memories associated with that temporal time-stamp feature, increasing the likelihood they will be selected for further processing. In contrast, as contextual demands decrease the probability for mindwandering increases, i.e., attentional selection expands beyond task-related representations (See Figure 4). The less contextual demands, i.e., more mindwandering state, the less bias selection in favor of particular representations. This leads to greater influence of bottom-up signals on internal selection such as the propensity to mindwander to certain content (e.g., negative spontaneous thoughts; (Marchetti, Koster, Klinger, & Alloy, 2016; Smallwood, 2013). Accordingly, the A2T contextual demand model component has two computational functions, to set (1) the context (i.e., task) relevant information, and (2) the degree of moderation of the influences of representations in WM and affect on internal attentional selection. Finally, it is notable that contextual demand has a strong tonic (Buetti & Lleras, 2016) effect on internal selection and thereby the contents of WM and affect. Indeed, relative to *representations in working memory*, its direction of influence on selection biases (i.e., degree of biasing selection in favor of task-relevant representations) is more consistent over longer periods of time. This is because unlike the representations component which decays quickly due to constraints of short-term memory (e.g., seconds), the contextual component is set by external factors may often persist over longer time periods (e.g., minutes).

## Model paths: Feedforward & Feedback

The components of the system are linked through a number of 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 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, these paths or processes are 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).

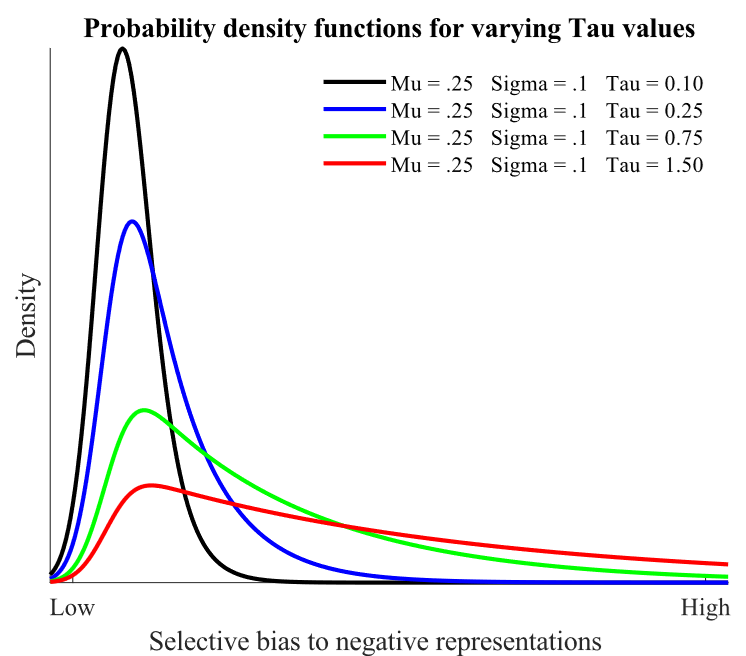
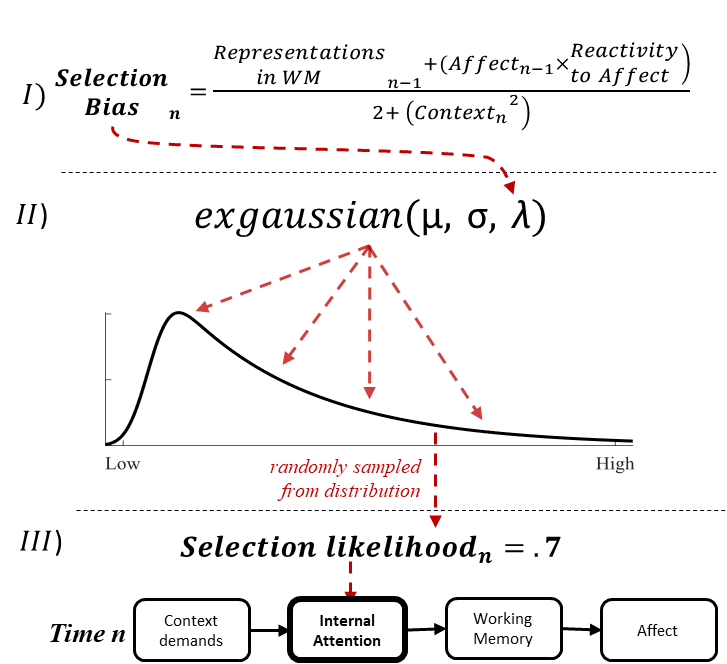
**First, *reactivity to representations in WM*** (path *c*) or “emotional reactivity” refers to the change in emotional state and its intensity (peak amplitude) in response to emotional information (Davidson, 1998). Reactivity to representations in WM is an important individual differences factor in 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 instructions such as “Think how you feel inside.” (Susan 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, *reactivity to affective state*** (path *e*) or “cognitive reactivity”, refers to the magnitude of increase in probability of negative cognitions (thoughts) in response to negative affect (Scher, Ingram, & Segal, 2005). In the model, this is reflected in the influence of affect on the internal attention selection likelihoods in favor of affect-congruent representations (Figure 1A Path *e*).

## **Heuristic formulation of momentary states of the system.**

As detailed above *internal attentional selection*is at the functional center of the A2T model and XXXXXX phenomena. Indeed, the momentary states of each component converge onto internal attentional selection, thereby systemically determining the state of the internal attention component at Time *n* likelihood of attending to a negative representation. To reiterate, although we here focus on negative XXXX representations, the model XXXXXXXX. Computationally, this likelihood is drawn from an exgaussian distribution whose parameters (mean, variance and exponential rate; see Table 2) 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, or more specifically, the degree (from 0 to 1) to which emotionally negative or neutral information is task-relevant[[4]](#footnote-4). For example, the mean parameter will be .5 when modelling an individual trying to recall a list with 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 rate (i.e., negative bias) increases, the distribution accordingly biases/shifts towards overall greater likelihood of attending to negative representations (see Figure 5).

|  |  |
| --- | --- |
| **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; λ) | Degree of bias to negative information. Integrates the states of representations in WM and affect (Time n-1) and contextual demands (Time n). As tau increases the distribution shifts more toward higher negatively biased selection likelihoods. |

****

**B**

**A**

**Figure 5.** **A) Illustration of change of ex-gaussian distribution as a function of tau values.** As tau increases (i.e., XXXX), the more the distribution is skewed to higher levels of selective bias to negative representations. **B)** Illustration of computational procedure of determining selection bias at Time *n*. In brief, first the states of contextual demands (Time *n*), representations in WM (Time *n-1*) and affect (Time *n-1*) are integrated using formula to determine the tau parameter (part I) together with the mean and variance (part II). Third. From this distribution function a single value is randomly sampled. This XXX value determines the selection likelihood at Time *n*. 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.

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 change from moment-to-moment since it is determined 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):

where *i* represents the individual person/model, *n* the specific moment in time. Accordingly, *selection bias* is a value between 0 (no bias to non-negative stimuli) to 5 (complete bias to negative representations) and is determined by values of the following elements: (a) the product of *affectn-1* and *reactivity to affective state*, and (b) the valence of *representations in WM n-1*; The average of these two elements is divided by (c) *Contextn*, the demand for focused attention at that moment in time. *Context* is squared and placed in the denominator to reflect the strong influence of contextual demands on internal attentional processes.

The degree of experienced negative *affect* is determined by 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):

after the selection likelihood distribution has been determined by the parameters (mean, variance, tau)asinglevaluethe exgaussian 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 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 is influenced by the momentary bias of the internal attentional selection component.

These heuristic equations help to illustrate how (a) the system state in context converges on internal attention and is aggregated into in a single value of Selection likelihoodn and (b) is in large part influenced by the values of *representations in WM* and *affect* at Time *n-1*; which itself is influenced by the probabilistic outcome of selection likelihoods of the *internal attention* component at Time *n-1*. Critically, this represents through the equation the temporal and dynamic nature of the model – that the components of the model inter- and trans- act from moment-to-moment in time (see Figure 1B). For example, while affect at Time *n* is determined by the representations in WM at that moment in time, affect also influences internal attention at Time *n+1* which, in turn, will determine representations in WM 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 *n*), a dynamic trajectory (i.e., a “greater whole”) emerges representing the unfolding of attention, thought and affect from moment-to-moment in time, as detailed below.

## Common temporal trajectories of the dynamic system

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*. The dynamic trajectories of the system are a function of the 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 from moment-to-moment 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 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.

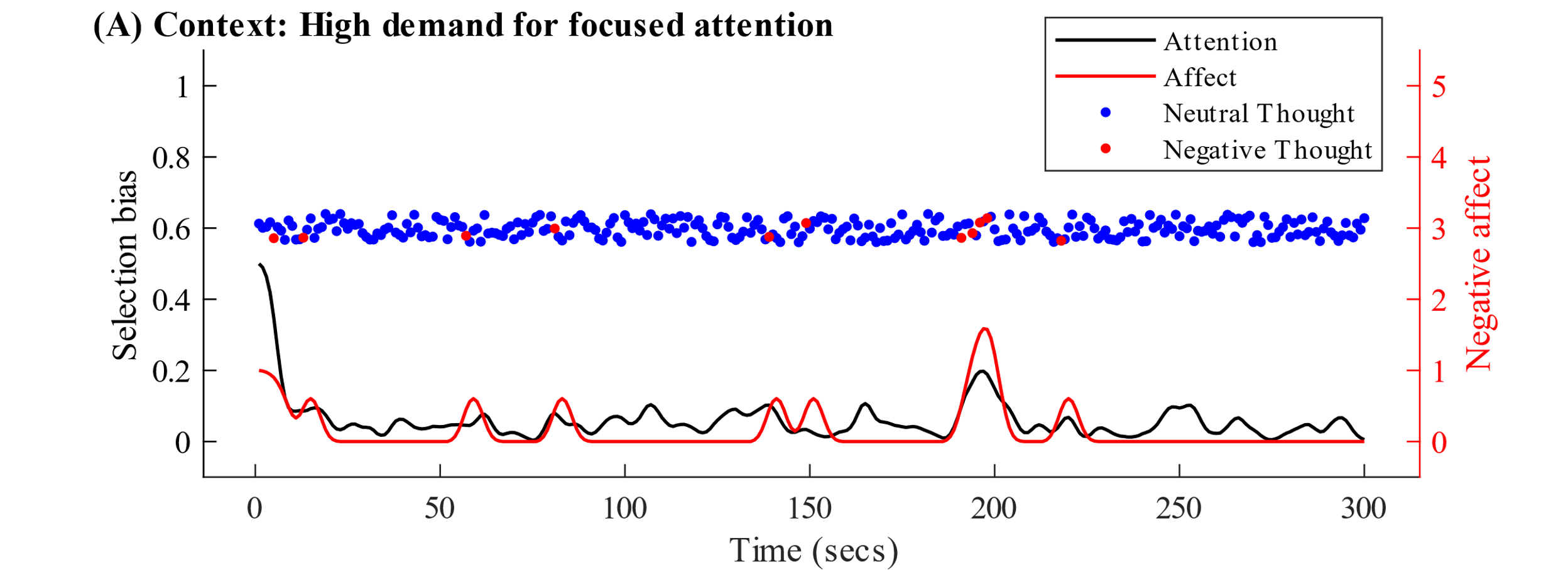
For example, trajectories may be characterized by greater stability or, conversely, variability in thought content (i.e., representations in WM) and related XXXXX over time. Stability is reflected in temporal sequences of thematically similar/related thought content (e.g., a series of task-related thoughts). 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 (CITE). 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. For example, depression has been conceptualized as an over-rigidity of thought (Christoff et al., 2016) whereas increased temporal variability in though-content is associated with high schizotypy trait (Welhaf et al., 2019).

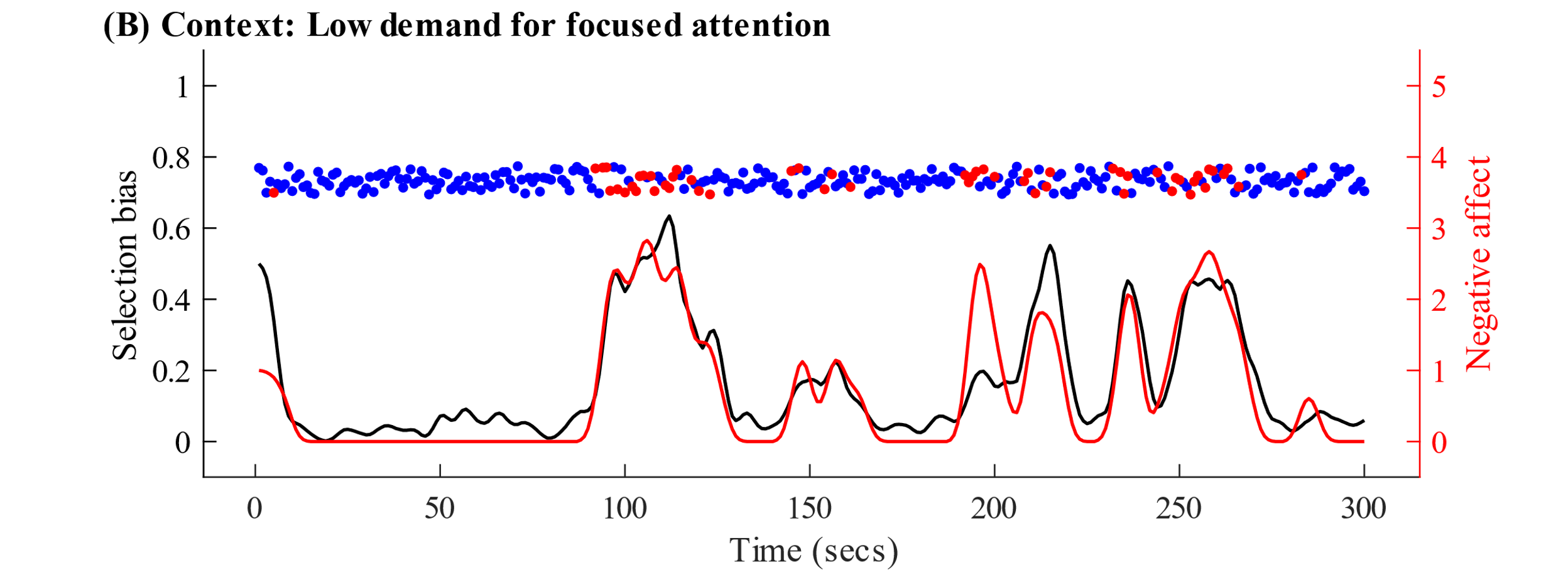
To illustrate how the model enables identifying, modelling and predicting the unfolding or emergence in time of different trajectories from an initial state in context, we begin by describing how high-demand context (or task-oriented states) versus low-demand context (or mind-wandering states) influence Internal Attentional Selection in determining likely trajectories of selection likelihoods and levels of negative affect, over repeated states of the system, over time. We then describe how these contextual demands and respective mindwandering/task-oriented states interact with components of representations in working-memory and affect to influence internal attentional selection and thus various probable trajectories. To do so, we describe and simulate the probable trajectories when emotionally neutral information is processed and when negatively-valenced information is processed.

### **High-demand context: task-oriented states.**

High-demand context to focus on neutral information. Contexts of high-demand to focus attention on specific goal- or task- relevant information 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 initiates and is maintained until contextual demands for sustained focused attention decreases. For example, trying to recall the author and name of a paper to cite requires that internal attention is constrained to associated information in long-term memory, until the required information is retrieved.

To illustrate how the model is likely to function under these conditions, we simulated the dynamic system over a 5-minute time period. See Supplementary Materials (SM) for details on the simulation algorithm and procedure. Figure 5A illustrates 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, due to the fact that selection likelihoods are normally distributed (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 attended. Accordingly, when this random selection occurs sequentitally (see Figure 5A ~190 second 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 this “random” or low probability events.



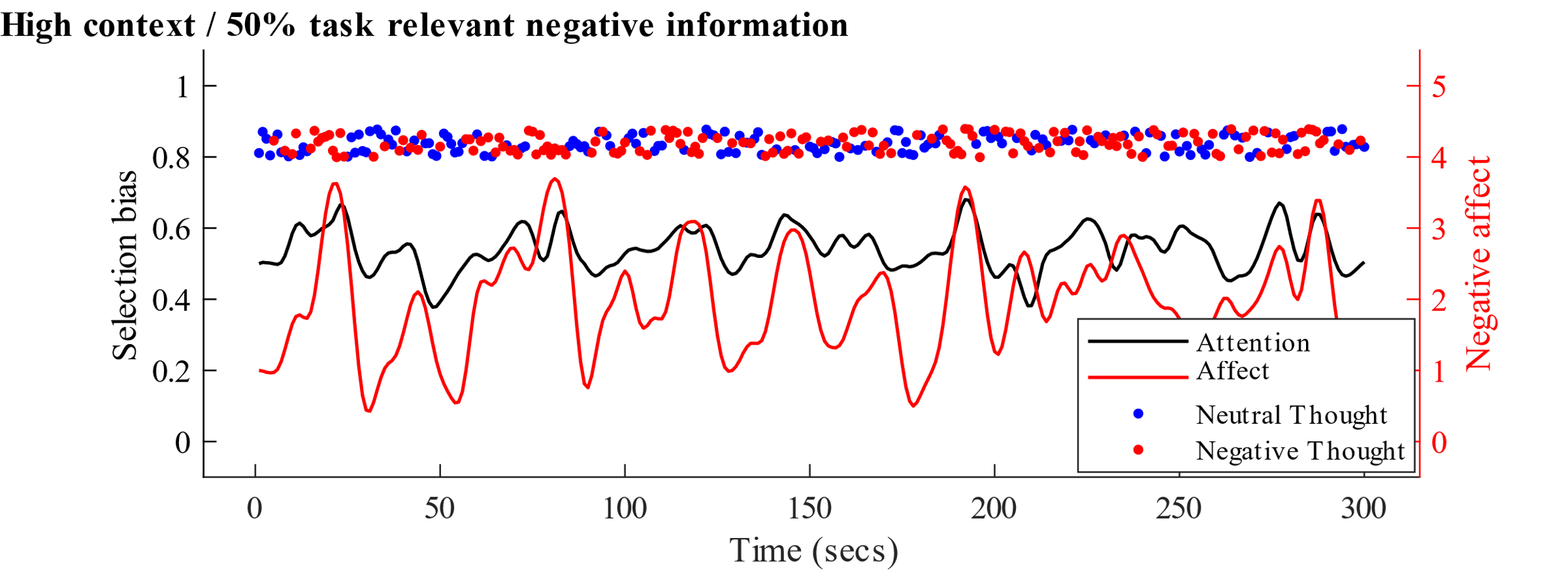


**Figure 6.** Simulated values in time (5 minutes) of Internal Attention and Affect components. Red line = degree of negative affect at each moment in time. Black line = selection bias – the likelihood of selectively attending to negative representations. Importantly, the selection bias is a *likelihood* *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 Figure 5A the contextual demands is set to high (= 4) and in 5B 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). See Supplement Materials for details of simulation algorithm.

However, context interacts with the other components which also influence internal selection likelihoods. For example, as illustrated in Figure 2, in addition to contextual demand, activated representations in WM bias subsequent selection in favor of representation-related content (see Figure 2). Therefore, if the active representation in WM is task-related then there are two components that bias selection in favor of task-related information: (1) the contextual demand for focused attention and (2) the active representations in WM. Consequently, task-oriented states should be temporally stable (as seen in Figure 6A) relative to other (more mindwandering) trajectories wherein context and WM representations are *not* necessarily congruent, driving greater variability in selection bias over time (as seen in Figure 6B; also see below).

High-demand context and stochastic nature of internal attentional selection. Moreover, due to the partially stochastic nature of the Internal Attentional Selection component – wherein selection is represented as a probability and so occasionally low probability events occur – there may be random, and relatively infrequent, occurrences wherein task-unrelated thoughts are selected. This is reflected visually in the random (yet infrequent) selection of negative task-unrelated representations (red dots) in Figure 6A. 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 tonic/sustained biasing effect of contextual demands in favor of task-related information (see Figure 1B).

High-demand context to focus on negatively valenced information. When task-related information is *negatively-valenced*, for example trying to recall important information from a traumatic experience/memory or to plan a stressful event, the 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 *both negative and neutral* information, such as memorizing a list of mixed negative and neutral words, 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, 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., 2019).



**Figure 7.** 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.

### **Low-demand context: mind-wandering states.**

Low-demand context to focus on neutral information. 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). In this context, the representations in WM, affect as well as stochastic element of internal attention have greater weight in influencing selection. The contents of WM influence selection such that current representations in WM (Time *n*) bias subsequent (Time *n*+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 *n-1*) is selected – switching the ‘theme’ of 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 visualised simulation).

Interactions of low-demand contextual demands and affect. As described above, affect also influences selection of internal content by biasing selection in favor of affect-congruent content. Therefore, when negative affect is triggered (Time *n*), it then biases subsequent (Time *n*+1) selection in favor of (negative) affect-congruent representations (e.g., negative thoughts about oneself). If a negative representation is selected into WM (Time *n+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 selection and so on (Whitmer & Gotlib, 2013). 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).

## Potential applications of the Attention-to-Thoughts model

A2T model provides a conceptual and computational framework for examining the temporal dynamics of internal attention, thought and affect. Accordingly, A2T has potential applications to the study of psychological phenomena or processes characterized by temporal dynamics of thought and affect. For example, A2T has direct and significant implications for understanding the temporal dynamics of cognitive-emotional processes in mental health. Specifically, A2T has important implications for understanding common forms of cognitive vulnerability that contribute to the onset and maintenance of prevalent forms of psychopathology.

Cognitive vulnerabilities is an umbrella term for cognitive processes, that when dysregulated, have been shown to predict common forms of mental health problems (Mathews & MacLeod, 2004). Collectively, these vulnerabilities are typically characterized as being *internal* (focus on thought content and thinking styles), *temporal* (how thought content and style is 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) (Fox et al., 2018; Harvey, Watkins, & Mansell, 2004; Koster, De Lissnyder, Derakshan, & De Raedt, 2011; Susan Nolen-Hoeksema et al., 2008; Siemer, 2005). As such, these key characteristics make cognitive vulnerabilities a good illustrative example of the application and utility of A2T in advancing theory.

*Difficulty to disengage*

*Captures mental capacity*

*Distraction*

*Struggle to let go of thoughts*

*Interference by negative content*

*(Negative) funneling effect*

Figure 8

# Internal attention and mental health

A number of well-established forms of higher-level cognitive processes, when dysregulated, function as cognitive vulnerabilities for a host of common mental health problems including depression and anxiety (see (Mathews & MacLeod, 2004) for review). These include, for example, repetitive negative thinking, rumination and worry (Borkovec, Robinson, Pruzinsky, & DePree, 1983; Ehring & Behar, n.d.; Ehring & Watkins, 2008; S 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) and impairments in cognitive control such as increased interference by negative 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). Critically, many of these vulnerabilities more often than not co-occur and interact with one another (Hong & Cheung, 2014; Mansell & McEvoy, 2017; Marchetti, Loeys, Alloy, & Koster, 2016). The commonco-occurrence of vulnerabilities and the respective mental health problems associated with them, we argue, may be related to a common role of internal attention. Specifically, we argue, first, that biases in lower-level internal attentional processes cut across and sub-serve these higher-level processes (see Figure 1 for illustration); and second, A2T may help to advance understanding of the emergence and maintenance of common forms of cognitive vulnerability. Among other implications, such insighs may be useful to to guide training or therapeutic targets to prevent and reduce vulnerability and thereby improve mental health.

There have been initial efforts to develop models of internal attention (or related cognitive processes) underlying cognitive vulnerabilities. Models include: (a) the attentional scope model of rumination, in which negative affect limits the scope of accessible content in long-term memory to mood-congruent information thereby increasing the likelihood of rumination (Whitmer & Gotlib, 2013); (b) the disengagement hypothesis model in which difficulty disengaging attention from self-referential information underlies ruminative states (Koster et al., 2011); (c) the cognitive inhibition model in which deficits in emotion regulation are driven by deficits in inhibiting selection and activation of no longer relevant negative information in WM (Joormann, 2010).

These (with perhaps the exception of attentional scope)e, and as such, A2T provides a framework for integrating both “top-down” effects (conceptualized as effects of contextual demands and representations in WM on internal attentional selection) with “bottom-up” effects (the effects of representations in WM and affect on selection) in driving maladaption and vulnerability.

Furthermore, the application of A2T to cognitive vulnerabilities expands on current models by providing a single, unified explanatory framework for the emergence of attentionally related maladaptive processes (e.g., scope, disengagement, inhibition), as opposed to a number of separate models/theories for each mechanism. Moreover, A2T provides added utility to current research by providing a computational basis for the phenomena of interest. That is, A2T not only *describes* what processes may become dysregulated but also provides a *computational* framework for how dysregulation emerges. This enables novel tools such as model simulations under different contextual and individual factors, resulting in specific quantitative predictions based on changes in these factors. For example, predicting the unfolding of thought content and emotional states as a function of individual factors (e.g., high trait reactivity to affect) and situational factors. Such simulations enable testing the overlap between simulated and empirically observed behavior (M. K. van Vugt & van der Velde, 2018). In summary, the application of A2T to cognitive vulnerabilities and mental health provides a novel conceptual and computational framework for understanding how typically adaptive low-level processes (of attention, WM and affect) inter- and trans- act in time and emerge as higher-order maladaptive processes (e.g., rumination and negative repetitive thinking).

## Higher-level processes: cognitive vulnerabilities and individual differences

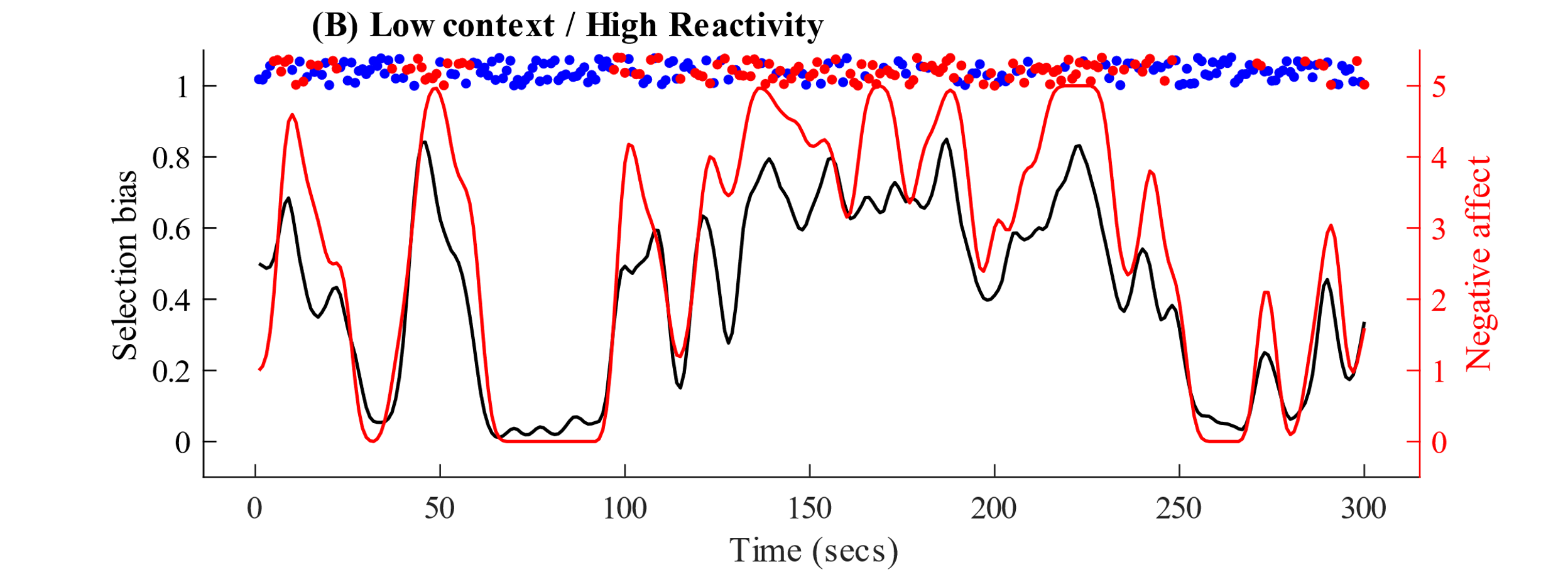
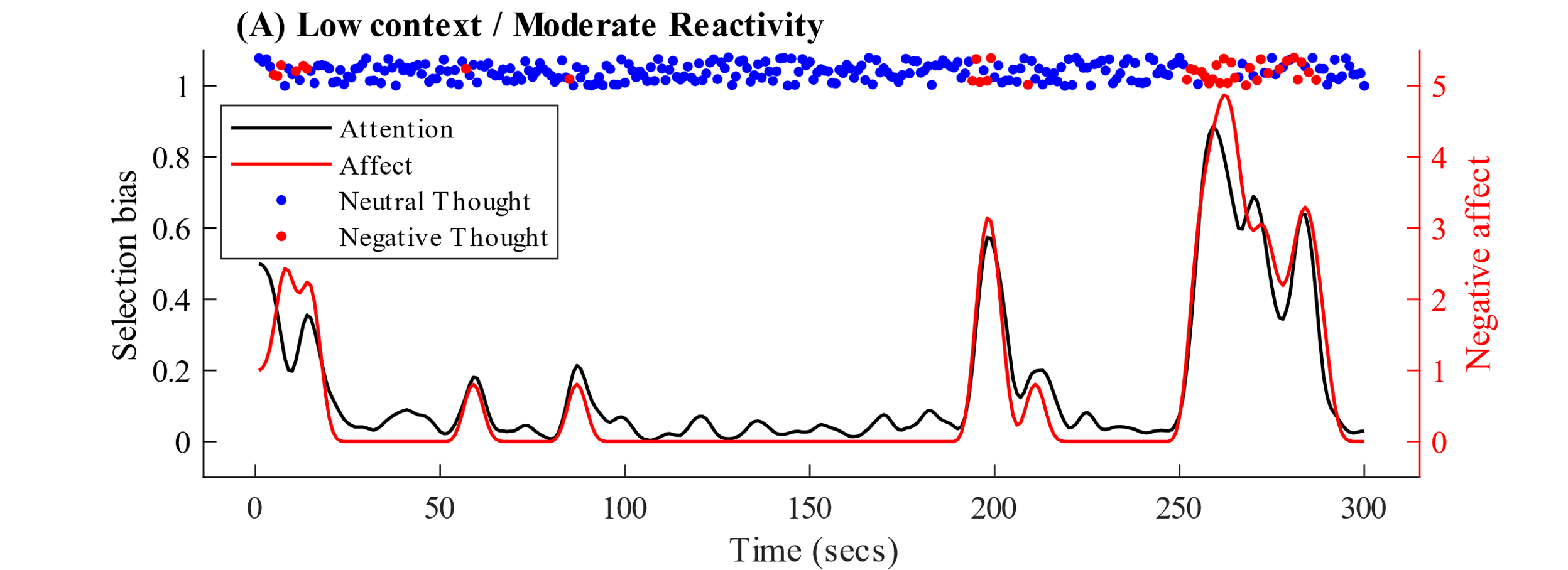
As described above, not only are momentary states of the dynamic system likely to potentiate adaptive trajectories of context-sensitive stability or variability in attentional selection in time (see Figure 6), but critically, when overly stable (i.e., rigid) dynamics of the system are chronic or likely to repeat over-time and context, then higher-level cognitive vulnerabilities and related mental health problems are likely to emerge as stable traits (Andrews-hanna et al., 2014; Watkins, 2008; Zabelina & Andrews-Hanna, 2016). In terms of the proposed A2T model, various common forms of cognitive vulnerability that characterize biased and overly rigid thinking, may simply be understood as chronic self-sustaining trajectories of the A2T system. To illustrate, we focus on three vulnerabilities that are well studied, implicated in mental health, and in which internal attention is theoretically implicated as a key mechanism – repetitive negative thinking, cognitive dyscontrol and interpretation bias (Ehring & Watkins, 2008; Everaert, Podina, & Koster, 2017; Grahek et al., 2018).

We focus on repetitive negative thinking as this phenomenon’s key characteristics especially align with descriptive “field” of A2T – selective processing of internal experiences (negative thought content) and temporality (*repetitive* thinking). We first highlight how low-level momentary states in context and trajectories of the model drive these higher-level cognitive vulnerability processes. We then relate to the role of individual differences in the model paths of reactivity to representations in WM and reactivity to affect in accounting for individual differences in the prevalence and severity of such vulnerabilities. Notably, beyond these particular forms of cognitive vulnerability for common mental health problems, A2T similarly accounts for XXXXXXX YYYYYY.

Repetitive negative thinking. Repetitive negative thinking (RNT) 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 & Watkins, 2008; Ehring et al., 2011). RNT, as well as related processes of rumination and worry, have been found to predict negative outcomes such as increased momentary negative emotions and ultimately the onset of depression as well as post-traumatic stress and generalized anxiety disorders (Hirsch et al., 2018; Watkins, 2008; Zetsche, Bürkner, & Schulze, 2018). RNT’s core characteristics of being repetitive (i.e., temporal) and capturing mental capacity (i.e., attentional) make it an ideal process for illustrating how A2T model explains the emergence of cognitive vulnerabilities.

Unfolding of RNT in A2T. A RNT state begins, typically under low contextual-demands (CITE), when a negative representation (thought) is selected into WM (CITE). This representation then triggers or increases negative affect (CITE). 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. At this point, it is helpful to again highlight a critical outcome of the stochastic element in internal attentional selection - enabling ‘random exits’ from RNT – meaning an instance (of random) selection of non-negative content breaking or attenuating the feedback loop (see Figure 8B time point ~240sec). Without a random element the model becomes overly rigid, determining that once a 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, whereas the proposed causal associations between contextual demands, representations in WM and internal attentional selection (paths *a* and *d*) are fixed, and thus not a source of individual differences in likelihood of certain maladptive trajectories to occur - the impact of representations in WM on affect (i.e., reactivity to representations in WM - path *c*) and the impact of affect on internal attentional selection (i.e., reactivity to affective state - path *e*) account for individual differences in maladaptive trajectory likelihoods. Critically, these two paths constitute the closed feedback loop from WM onto affect and then back onto internal attentional selection. Therefore, individuals with higher levels of reactivity have a stronger feedback loop, increasing the likelihood of 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. As can be seen, higher reactivity levels increase both (a) the frequency of entering- and (b) average duration of states of repeated selection of negative content into WM and related increase in negative affect.



**RNT**

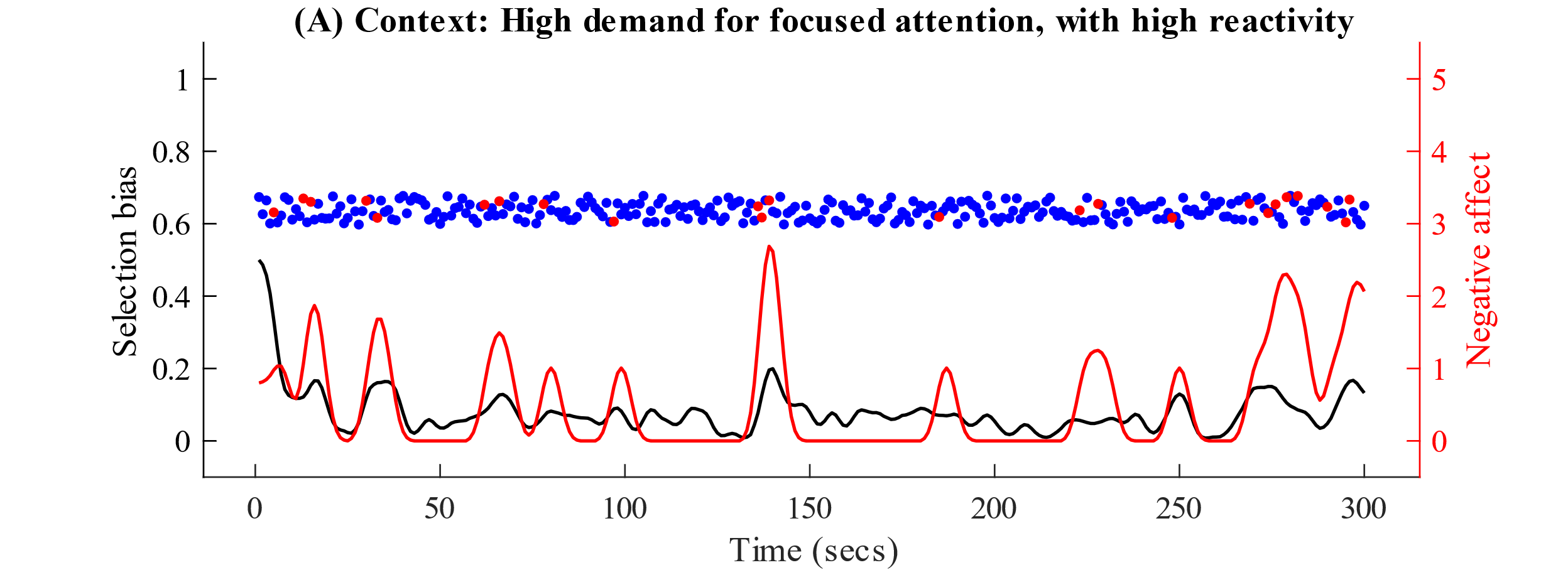
**Figure 7.** Simulated differences between cognitive and emotional reactivity level (Moderate = 0, High = .25). Context is set to low (= 2). All other initial component values are equal to those used for Figure 5A. RNT = Repetitive negative thinking.

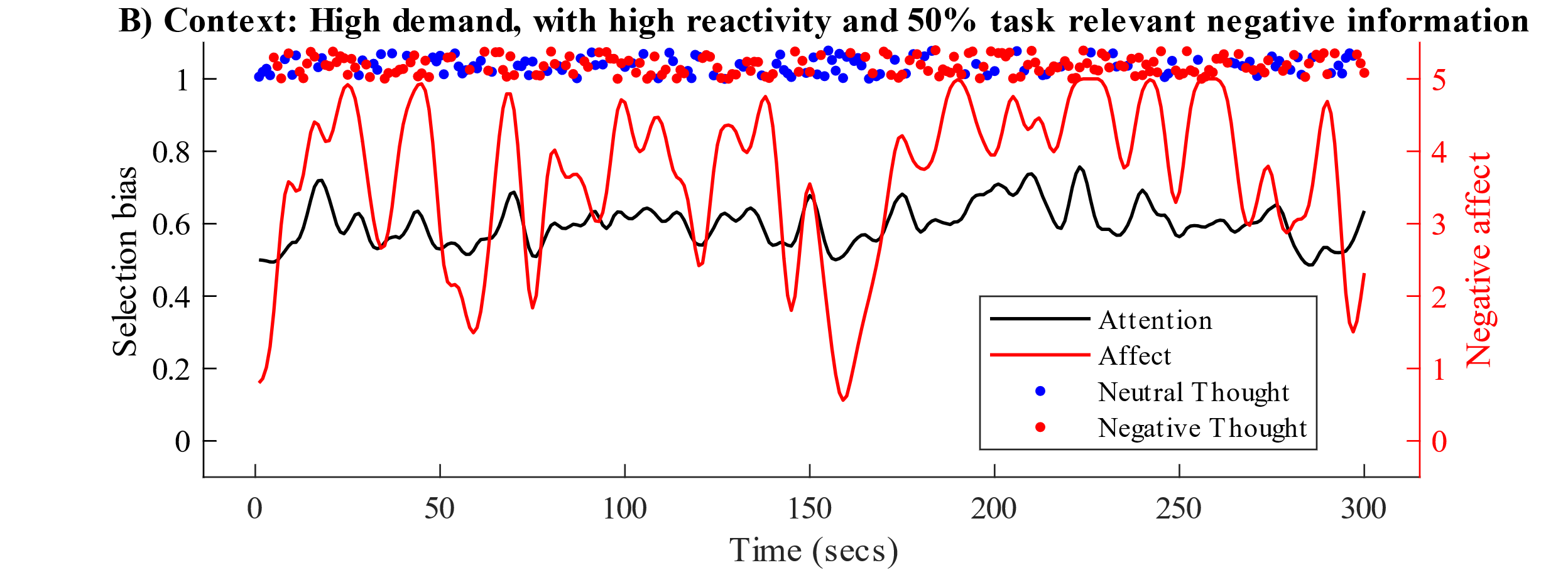
Cognitive (dys)control. Cognitive control refers to the ability to flexibly adapt cognition and behavior to current goals (Grahek et al., 2018), often used interchangeably with “executive control” or “executive functions”. Cognitive (dys)control accordingly covers various cognitive abilities including inhibiting prepotant 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 or domain-general (i.e., neutral) information (Grahek et al., 2018; Zetsche et al., 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. More precisely, we conceptualize cognitive control as a dynamic behavior which 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 the model we conceptualize cognitive control as the moment-to-moment competition (or differential influences) between the effects of contextual demands vs. the effects of representations in WM and affect components 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). 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.

The influence of contextual demands, 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.** 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 “homonculus” control component (Baddeley, 2011; Everaert, Grahek, & Koster, 2017). The proposed A2T model enables examining the stability of cognitive control (or conversely failures) as frequency of momentary states of attention directed towards/away from task irrelevant information – as a function of the different states in context and individual differences of low and high reactivity.

Individual differences in cognitive (dys)control. As described above, in addition to context, momentary changes in selection bias likelihoods are also determined by affect, representations in WM and reactivity paths, as well as the stochastic elements of internal selection. Figure 9A illustrates how a simulated model with high reactivity has random increases or spikes in selection likelihood of bias to negative task-unrelated representations, even under high contextual demands. 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 selecting *affect-congruent* *but potentially task-irrelevant* information. This likely leads to greater frequency of cognitive control failures in the form of attending to task unrelated thoughts. Figure 8B illustrates how high reactivity, together with negatively-valenced task-relevant information, increases the selection bias to negative representations (as the context requires) but has a stronger impact on variability and mean negative affect. As such, individual differences in emotional and reactivity to affect state 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; Schweizer et al., 2019). Moreover, these A2T simulations suggest that, under high contextual demands, momentary affective responses may be a more meaningful (i.e., have more variability) measure in response to manipulations of task-relevance/irrelevance of affective information than measures of only task performance (as a proxy of cognitive control) (Schweizer et al., 2019). Even more in-line with the A2T approach, it may be more conceptually accurate to make hypotheses and empirical tests on the momentary temporal level (rather than over-time aggregation) – 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; Smallwood & Schooler, 2015).





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

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; (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, et al., 2017; Hirsch et al., 2018). Interpretation bias’s core characteristic of preferential selection of one interpretation over competing resolutions to an ambiguous cue (Gernsbacher & Faust, 1991) makes it an additional ideal process to demonstrate how momentary states of the proposed A2T dynamic system are likely to potentiate (mal)adaptive trajectories and thereby higher-level cognitive vulnerability.

Interpretation bias in the model. 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. Which interpretation (negative, positive or neutral) is selected is determined by the selection likelihoods (at Time *n*) influenced by the immediately preceding state of the system – such as the active representations in WM and affective state (at Time *n-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 m then increase the likelihood that negative interpretation is selected for further processing (Aue & Okon-Singer, 2015), i.e., entering WM and reaching awareness. That is, people with *active* negative expectations/beliefs are more likely to demonstrate negative interpretation biases (Hutchinson & Turk-Browne, 2012). This illustrates how the model, and especially the component of internal attentional selection, interacts with other cognitive vulnerabilities (such as negative schemas about the world, future and oneself)[[5]](#footnote-5) to potentiate maladaptive behaviors and thinking processes. 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 the Figure 1A) and affect as outcomes rather causes of biased interpretations ((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 trans-action, that unfolds from moment-to-moment in time (see Fig 1B), between attention, WM and affect – i.e., they are reciprocally causes- *and* outcomes- of one another in time.

AMIT – PLEASE SKIP THE EMOTION REGULATION SUB-CHAPTER BELOW AND CONTINUE ON TO MINDFULNESS CHAPTER. I THINK WE SHOULD DROP IT. I FEEL IT DOESNT FIT IN WITH THE REST OF THE PAPER NOW. WE SHOULD DISUCSS.

Emotion regulation: distraction and reappraisal. Emotion regulation entails the various strategies or processes through which individuals attempt to modulate their emotional states to appropriately respond to environmental demands. Due to its regulatory and adaptive functions (Gross, 2007) deficits in the ability to choose and implement effective emotion regulation strategies contribute to vulnerability across common mental health problems (Aldao, Nolen-Hoeksema, & Schweizer, 2010). Emotion regulation is theorized as central to depression and anxiety which are often viewed/conceptualized as disorders of impaired emotional regulation (Aldao et al., 2010; Joormann & Gotlib, 2010; Mennin & Fresco, 2013). Emotion regulation is a complex multi-process higher-level phenomenon emerging from the interactions of lower-order processes such as emotion identification, attention and response selection (Sheppes et al., 2015). Here we focus on what we view as the *attentional* processes related to regulation and illustrate how the model can account for the most two commonly studied forms of emotion regulation – distraction and reappraisal (Thiruchselvam, Blechert, Sheppes, Rydstrom, & Gross, 2011).

Emotion regulation in the model. One common strategy for emotion regulation is distraction – the process of diverting attentional focus away from the emotionally salient and eliciting aspects of an event. In the framework of our model, and similar to all emotion regulation processes, a distraction process initiates when representations from a negative event (whether perceived or recalled from memory) enter WM and trigger negative affect that is inconsistent with a desired state (Carver & Scheier, 1990). In such a state, contextual demands for focused attention (towards distracting information) conflict with the Representations in WM and Affect on influencing/biasing selection likelihoods – i.e., a conflict between top-down (contextual goal-directed demands) and bottom-up (affect- congruent directing of attention) signals. Computationally, the mean (mu) parameter pulls the selection likelihood distribution towards preferential selection of non-negative information, while the exponential rate (tau, i.e., selection bias) pulls towards preferential selection of negative information. With each subsequent momentary state in context in which *distracting* information is selected it temporarily *reduces* the conflict between contextual and WM representation and affect influences Accordingly, distraction (i.e., successful emotion regulation) is achieved over time once enough non-negative/distracting representations are selected, enter WM and until sufficient reduction in discrepancy between current and desired emotional state is achieved. Distraction fails if the contextual demands signal cannot overcome competing and conflicting bias signals. This proposed process is in-line with prominent views of attention as the dominant mechanism underlying distraction strategy (Sheppes et al., 2015), although here we further elaborate on how the attentional process may unfold in time.

A second common strategy is reappraisal – the re-evalution of an event’s emotional meaning. Similar to distraction, a reappraisal process begins when representations from a negative event enter WM, elicit a negative affect discrepancy with desired affect state and a regulation goal is set to reduce discrepancy. Here it is important to restate that WM representations bias in favor of *content-related*- and Affect biases *affect-related*- information. This is important because in the reappraisal process, attention is maintained on event-related information (e.g., a negative interpretation). Thus, the proposed model predicts that the contextual demands for sustained focused attention and the representations in WM bias in favor of (somewhat) similar information whereas the effects of affect on internal attentional selection remain in conflict with contextual demands and WM representations. Successful reappraisal is achieved once enough iterations of the system occur with selection of information (representations) inconsistent with the undesired interpretation. Since these representations consume limited capacity WM resources they eventually ‘inhibit’ or remove the undesired appraisal from WM as well as its subsequent (undesired) emotional reaction. For example, when recalling an important job interview, a person might initially experience negative responses to certain parts of the interview, and in order to reappraise, tries to emphasize and recall specific positive elements of the memory. This means attention is maintained on the same general thematic content (the memory) but attention is also directed to specific positive information in the memory, until the undesired negative emotional state is reduced.

This attention-based approach to reappraisal differs from dominant conceptualizations which view reappraisal as a separable process occurring after attention has been allocated (Schönfelder, Kanske, Heissler, & Wessa, 2013), although reappraisal has also been conceptualized as an attentional process (Thiruchselvam et al., 2011; van Reekum et al., 2007). Here, we highlighted how successful reappraisal may, at least in part, be driven by internal attention as it requires a certain degree of sustained attention on (i.e., repeated selection of) specific internally represented information.

One key prediction from this model is that conflict begins at an earlier stage in distraction (as soon as representations in WM begin to exert influence which conflicts with contextual demands influence) relative to appraisal (as soon as emotion begins to exert influence). This prediction is in line with EEG findings showing that distraction modulates event-related potentials prior to appraisal (Schönfelder et al., 2013; Thiruchselvam et al., 2011).

Individual differences in (deficits in) emotion regulation. We view reactivity to representations in WM and reactivity to affect as the main factors explaining individual differences in the (in)ability in emotion regulation. As describe above, both strategies function through altering attentional selection. Since persons with higher levels of reactivity to representations in WM have higher negative affect responses to negative representations, together with higher reactivity to affective state accordingly having a stronger influences of affect state on attentional selection, such individuals should require more effort, resources or time for contextual demands (and in appraisal) representations in WM to overcome the effects of affect on selection.

## Future directions

The A2T model reflects initial efforts to understand the role of internal attention in thinking and feeling. Future work could focus on a number of key questions and avenues of research. First, A2T’s computational framework enables simulating specific contexts and in individual differences. From these simulations, specific quantitative predictions can be made (e.g., frequency of off-task thoughts during a certain time-period, affective levels as a function of representations in WM and vice-versa) and can be tested with empirical data. XXXXXX.

Second, many aspects of the model can be further specified and examined. For example, (a) what is the source of the internal objects (e.g., how do spontaneous thoughts arise?; (Axelrod et al., 2017; Smallwood, 2013)) and (b) how many different objects compete for attentional selection? The issue of multiple competing objects is similar to issues for example in visual attention. The visual world is (almost infinitely) rich in detail but for controlled experimental research this richness is reduced into a controlled (typically small) number of competing visual objects. Future research should aim further develop both theoretical models and methodology for experimentally controlling objects competing for selection (e.g., Amir et al, in press). WhilY THAT MAY BE e WM research has extensive methods, it is often difficult to dissociate attentional and memory processes (Baddeley, 2011).

## Limitations

Despite its potential utilities, the model currently has several key limitations. First, to provide a broad account, certain components are characterized using abstract terms. For example, the term “context” here implies both external context (e.g., external situational factors such as social demands) and internal contexts (e.g., top-down goals such as regulating one’s emotion). Second, for clarity and brevity, we characterized the different components in mostly categorical terms – for example, high or low contextual demands. However, the various components and their influencing paths can be dimensional. For example, XXXXXXX. Accordingly, in extensionso and applications of A2t, XXXXX. Third, the model is based on competition between various top-down, and critically, a number of bottom-up signals (e.g., different spontaneous thoughts and associative memories competing for awareness). However, the number of different bottom-up signals is unknown. How many different objects compete for attentional selection at any given moment? This problem is similar to issues for example in visual attention. The visual world is (almost infinitely) rich in detail but for controlled experimental research this richness is reduced into a controlled (typically small) number of competing visual objects. Fourth, there are currently limited experimental methods for studying internal attentional processes. Clearly, work is needed to develop rigorous methodologies. An additional path forward is to study ‘meta’-parameters. For example, simulation studies can approximate the probability/ frequency and duration of mindwandering episodes as a function of context demands and individual differences in reactivity. Such parameters (e.g., frequency of episodes over 20 minute simulation) can be compared to intensive experience sampling data (e.g., thought probes over a period of 20 minute) (Smallwood & Schooler, 2015; Welhaf et al., 2019). Finally, the A2T model can be further developed and expanded. For example, individual differences in WM capacity can be incorporated to simulate and test their salutary and/or maladaptive effects on trajectories of the system (Welhaf et al., 2019). Furthermore, A2T can be specifically “tailored” for specific phenomena of interest. For example, researchers interested in meditation practice can tailor the model to simulate the competition between goal-directed attention to mediation instructions (e.g., focusing on breath”) and competing spontaneous thoughts.

1. We capitalize component names (e.g. ***A***ffect) to separate the name from the common meaning of the term (e.g., *affect* as a momentary negative or positively valenced emotional experience). [↑](#footnote-ref-1)
2. However, the dimension on which selection is based can be changed when applying the model to specific topics of interest. [↑](#footnote-ref-2)
3. For computational/simulation purposes we view this component as a repeatedly updating first-in-first-out storage device containing the last 5 selected representations/objects. [↑](#footnote-ref-3)
4. Or any other feature/dimension of thought (e.g., self-referentiality) of interest, depending on what phenomena of interest A2T is applied to. [↑](#footnote-ref-4)
5. When such *higher-level* processes are active they trigger *lower-level* processes, such as representations (for example negative beliefs in the form of verbal thoughts) in WM. As such, different higher-level processes interact through their shared underlying lower-level processes. [↑](#footnote-ref-5)