AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Psychopathology and Clinical Science

ISSN: 2769-7541

https://doi.org/10.1037/abn0000972

Toward Quantitative Cognitive–Behavioral Modeling of Psychopathology: An Active Inference Account of Social Anxiety Disorder

Yinghao Zhang[1], Friederike Elisabeth Hedley[2], Ru-Yuan Zhang[3, 4], and Jingwen Jin[1, 5]

[1] Department of Psychology, The University of Hong Kong

[2] Department of Psychology, University of Cambridge

[3] Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology

[4] Key Laboratory of Brain–Machine Intelligence for Information Behavior—Ministry of Education, Shanghai International Studies University

[5] State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong

Understanding psychopathological mechanisms is a central goal in clinical science. While existing theories have demonstrated high research and clinical utility, they have provided limited quantitative explanations of mechanisms. Previous computational modeling studies have primarily focused on isolated factors, posing challenges for advancing clinical theories holistically. To address this gap and leverage the strengths of clinical theories and computational modeling in a synergetic manner, it is crucial to develop quantitative models that integrate major factors proposed by comprehensive theoretical models. In this study, using social anxiety disorder (SAD) as an example, we present a novel approach to formalize conceptual models by combining cognitive–behavioral theory (CBT) with active inference modeling, an innovative computational approach that elucidates human cognition and action. This CBT-informed active inference model integrates multiple mechanistic factors of SAD in a quantitative manner. Through a series of simulations, we systematically examined the effects of these factors on the belief about social threat and tendency of engaging in safety behaviors. The resultant model inherits the conceptual comprehensiveness of CBT and the quantitative rigor of active inference modeling, delineating previously elusive pathogenetic pathways and enabling the formulation of concrete model predictions for future research. Overall, this research presents a novel quantitative model of SAD that unifies major mechanistic factors proposed by CBT and active inference modeling. It highlights the feasibility and potential of integrating clinical theory and computational modeling to advance our understanding of psychopathology.

General Scientific Summary

Decades of research have yielded a rich array of theoretical models concerning the development and perpetuation of pathological behaviors. Recent progress in computational modeling has introduced tools for quantifying the complex interplay of factors giving rise to observable behaviors. Leveraging an established cognitive–behavioral theory (CBT) of social anxiety disorder, and employing the active inference modeling approach, this study serves as a demonstration of constructing CBT-informed quantitative models.

Alexander J. Shackman served as action editor.

Jingwen Jin https://orcid.org/0000-0001-9227-6837

Parts of this research were presented at the Annual Meeting of Society for Research in Psychopathology (September 2023). This work was supported by the National Key R&D Program of China (2023YFF1204200), National Natural Science Foundation of China (32441102 and 32100901), the Open Research Fund of the State Key Laboratory of Brain-Machine Intelligence Zhejiang University (BMI2400021), and Key Laboratory of Brain-Machine Intelligence for Information Behavior-Ministry of Education (2023JYBKFKT005) to Ru-Yuan Zhang. The authors would like to acknowledge the contribution of the preprint community. The authors thank Emmett Larsen for his helpful feedback. The authors declare that they have no conflicts of interest. This study was not preregistered. The present study contains no empirical data because the quantitative outcomes are all based on simulations. The model construction and simulations in the current study were implemented using standard routines (here spm_MDP_ VB_X.m) available as Matlab code in the latest version of Statistical Parametric Mapping (SPM) academic software (https://www.fil.ion.ucl.ac.uk/spm/). All simulations in the current study can be replicated (and customized) by running (and adjusting) the Matlab code included in the online supplemental materials. This research has been previously shared as a preprint available at: https://www.researchgate.net/publication/379427922_Toward_quantitative_cognitive-behavioral_models_of_psychopathology_An_active_inference_account_of_social_anxiety_disorder.

Yinghao Zhang served as lead for formal analysis and writing–original draft and contributed equally to conceptualization. Friederike Elisabeth Hedley served in a supporting role for writing–review and editing. Ru-Yuan Zhang served in a supporting role for writing–review and editing. Jingwen Jin served as lead for supervision. Yinghao Zhang and Jingwen Jin contributed equally to writing–review and editing.

Correspondence concerning this article should be addressed to Jingwen Jin, Department of Psychology, The University of Hong Kong, Room 661, 6/F, The Jockey Club Tower, Centennial Campus, Pokfulam Road, Hong Kong SAR, China, or Ru-Yuan Zhang, Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology, 112 Zhongyuan, 1954 Huashan Road, Xuhui District, Shanghai 200030, China. Email: frances.jin@hku.hk or ruyuanzhang@sjtu.edu.cn

Keywords: social anxiety disorder, active inference, computational modeling, Bayesian modeling, generative model

Supplemental materials: https://doi.org/10.1037/abn0000972.supp

Conceptual models of psychopathology have undergone rapid advancements in the past decades (Eaton et al., 2023; Kent et al., 2023). In contrast, the progress toward quantitative models of psychopathology, specifically those capturing the mechanisms and trajectories of the development of maladaptive behaviors, has been slow (Haslbeck et al., 2021; Jin et al., 2023). This discrepancy poses a crucial barrier to the mechanistic understanding of psychiatric conditions (Friston, 2023; Wang & Krystal, 2014). Conceptual models commonly employed in psychotherapy, while being highly valuable in guiding clinical conceptualization and treatment, are often criticized for being vague and difficult to falsify (Eronen & Bringmann, 2021; Meehl, 1978). Relatedly, formulating effective personalized treatments has proven to be a challenge and heavily relies on clinicians' intuition (Schneider et al., 2015; Wright & Woods, 2020).

Different from conventional conceptual models, quantitative modeling of hypothesized symptom-generation processes can provide invaluable insights by unambiguously and formally delineating vulnerability factors, their mechanistic relationships, and the pathogenesis trajectory through mathematical means (Huys et al., 2021; Montague et al., 2012; Stephan & Mathys, 2014; Wang & Krystal, 2014). Such models thus enable precise measurement and quantification of mechanistic factors and generate specific predictions of psychopathology profiles, holding the promise of facilitating personalized treatment strategies. However, in stark contrast to the comprehensiveness of conceptual models, the application of quantitative modeling to investigate the symptom generation process so far has primarily focused on isolated factors as detailed below. Since psychological disorders typically arise from a complex interplay of multiple factors, the feasibility and value of quantitative modeling in advancing our understanding of psychopathology depend on their ability to formalize the complex dynamics of these factors.

Social anxiety disorder[1] (SAD) is an example of a multifactorial psychological disorder characterized by excessive fear of negative social evaluations and avoidant behaviors (American Psychiatric Association, 2022; Leary & Kowalski, 1997; Stein et al., 2017). It is associated with significant functional impairments (Alonso et al., 2004; Leon et al., 1995) and economic burden globally (Fink et al., 2009; Grant et al., 2005; Javaid et al., 2023). While there exist well-established psychological models of SAD, its pathological mechanisms remain elusive, making it an important area of research. In the present investigation, we aim to elucidate pathways of SAD symptoms by leveraging the strengths of cognitive–behavioral theory (CBT) as the conceptual framework and active inference as the quantitative framework. We chose to employ these two frameworks because of their shared emphasis on the dynamics between cognition and action (Biehl et al., 2018; Hauke, 2016).

Under the CBT framework, individuals interact with their environment based on their beliefs or schemas about the self, the future, and the world, which are shaped by nature and nurture (Beck, 2021; Craske, 2017; Powers et al., 2017). These beliefs guide the selection of behaviors to reduce distress and obtain rewards. Consequences of these behaviors serve as feedback to update (including strengthening) the individual's beliefs. Psychopathology arises when distorted beliefs and maladaptive behaviors are present in a given environment (Beck, 2021; Craske, 2017; Mace et al., 1991). Psychotherapies based on CBT have been widely successful in treating various mental health conditions (Cuijpers, 2019; Hofmann et al., 2012). CBT provides a holistic conceptualization of SAD and effectively guides its interventions (Clark & Wells, 1995; Hofmann, 2007; Roth & Heimberg, 2001). Several classic CBT models, such as the models proposed in Clark and Wells (1995), Rapee and Heimberg (1997), and Hofmann (2007), have gained empirical support and have been widely used in clinical settings.

In the present study, we use the Hofmann (2007) model as the target conceptual model to demonstrate the suitability of constructing a comprehensive CBT-informed quantitative model. Building on its predecessors, this relatively recent model includes key components of earlier models and takes into account both broad as well as disorder-specific factors, resulting in a comprehensive CBT model of SAD (Hofmann, 2007). At the core of this model is the negatively biased belief that social situations are evaluative and threatening (social apprehension). This belief is associated with low self-efficacy (inadequate social competency) and catastrophizing about negative social outcomes. Such cognitions lead to safety (avoidance) behaviors, which in turn prevent individuals from gaining corrective experiences. Importantly, the model also highlights the cognitive style of self-focused attention, where individuals rely more on internal bodily sensations (interoceptive) than external environmental (exteroceptive) cues to evaluate the level of threat in social situations, often leading to inaccurate interpretations and a disconnection from the social environment (Hofmann, 2007; Spurr & Stopa, 2002). Additionally, rumination, particularly its maladaptive component, is characterized by dwelling on negative aspects of events and can reinforce negative social experiences as well as perpetuate negative social beliefs (Joormann et al., 2006; Nolen-Hoeksema & Morrow, 1991). The dynamic between these cognitive factors and behavioral avoidance is proposed to maintain social anxiety.

Although the Hofmann model (2007) is clinically useful and has gained empirical support in its efficacy (Andino et al., 2024; Carlbring et al., 2012; Enock et al., 2014; Hofmann & Otto, 2008), it lacks specificity and quantitative rigor in explaining how exactly different factors contribute to the development and maintenance of distorted social beliefs and avoidance behaviors. For instance, it remains unclear whether all the aforementioned factors have an equal influence on negative social beliefs and avoidance behaviors. Relatedly, the precise mechanisms through which CBT interventions produce therapeutic effects are often elusive (Butler et al., 2006). Additionally, such conceptual models often take a generic approach, leaving the heterogeneity of SAD underspecified (Binelli et al., 2015; Kopala-Sibley et al., 2014; Woods et al., 2020). The general

[1] In the present study, we use the term SAD to describe the general phenomenon, which encompasses symptoms that cause significant distress to the individual. We do not place emphasis on the categorical diagnosis, as proposed in taxonomic and diagnostic tools, which are based on symptom count and specific cutoff.

limitations of current CBT models are reflected in the Hofmann model of SAD. These common limitations hinder their utilization in research and clinical settings. Developing quantitative versions of these conceptual models that allow for the modeling of intraindividual processes of change can help unlock the full potential of CBT.

Active inference is a quantitative modeling approach that can provide precise and predictive insights beyond conceptual models. It is a Bayesian account of perception and action in which beliefs and behaviors are bridged in a manner similar to CBT (Friston et al., 2017). The theoretical alignment between active inference and CBT makes active inference a naturally suitable candidate for formalizing CBT-based conceptual models.[2] According to the active inference framework, the human mind acts like a scientist who constantly models the interactions between the self and the environment (Friston, 2010). This model is referred to as a generative model, which is essentially a belief system encoding how the environment (including the body) works by predicting what actions are likely to produce what (observable sensory) outcomes. These predictions guide the selection of actions to obtain desirable outcomes and avoid undesirable ones. This is consistent with an essential idea of CBT that schemas or belief systems guide behaviors. Actions are executed to impact the environment, which yields new observable sensory outcomes that influence belief updates (Friston et al., 2017; Parr et al., 2022; Sajid et al., 2021). This is consistent with how the consequences of behaviors drive changes in beliefs in CBT. Because we can only infer the causes of our observations and never know the hidden truth, this generative model can be imperfect, leaving room for inaccurate beliefs about the environment and predictions of actions. In addition, because the generative model also encodes how we perceive the outcomes of actions, not all outcomes generated by the environment are faithfully captured; instead, perception can be biased. This is consistent with cognitive and perceptual biases in CBT.

Active inference has been well-formalized (Parr et al., 2022) and applied to explain various cognitive functions, such as attention (Holmes et al., 2021), learning (Friston et al., 2016), interoception (M. Allen et al., 2022), as well as social conformity and decision making (Constant et al., 2019). Beyond basic sciences, it has been used to understand specific symptoms in psychiatric disorders, such as psychosis (Limongi et al., 2018), compulsive behaviors (Kiverstein et al., 2019), and aversive responses to auditory stimuli in posttraumatic stress disorder (Linson & Friston, 2019). These existing studies, however, have mainly focused on isolated factors or specific symptoms, with scant efforts to formalize unified and clinically relevant conceptual models (Haslbeck et al., 2021; Wright et al., 2023). For example, a study that explored the combination of active inference and CBT (Smith et al., 2021) focused on understanding how CBT can modify the approach-avoidance behavior. While this study yields important insight into a key aspect of anxiety, CBT models for specific disorders (e.g., social anxiety as used in the present study) typically include multiple interrelated mechanistic factors, many of which are specific to the disorder. Such complexity is often lost in quantitative research focusing on an isolated component.

In the present study, our overarching aim is to demonstrate the suitability of integrating CBT conceptualization and active inference modeling to develop comprehensive and quantitative models. For this purpose, our first aim was to quantitatively formulate the Hofmann (2007) model using active inference (Figure 1). Then, through eight simulations, we used the newly constructed quantitative CBT model to delineate differential pathways influencing social belief and avoidance behavior. Given that the present article combines both clinical theory descriptions and technical components from active inference, to ease the reading, we provide three levels of technical details: Audience interested in what can be achieved in our approach can skip the Methodology section and find a summary in the first section of the Results section; audience interested in the procedure of the model construction can find sufficient information in the Methodology section; audience interested in the technical details, including the specific code, can find the corresponding information in the online supplemental materials.

Methodology

A Brief Introduction to Active Inference

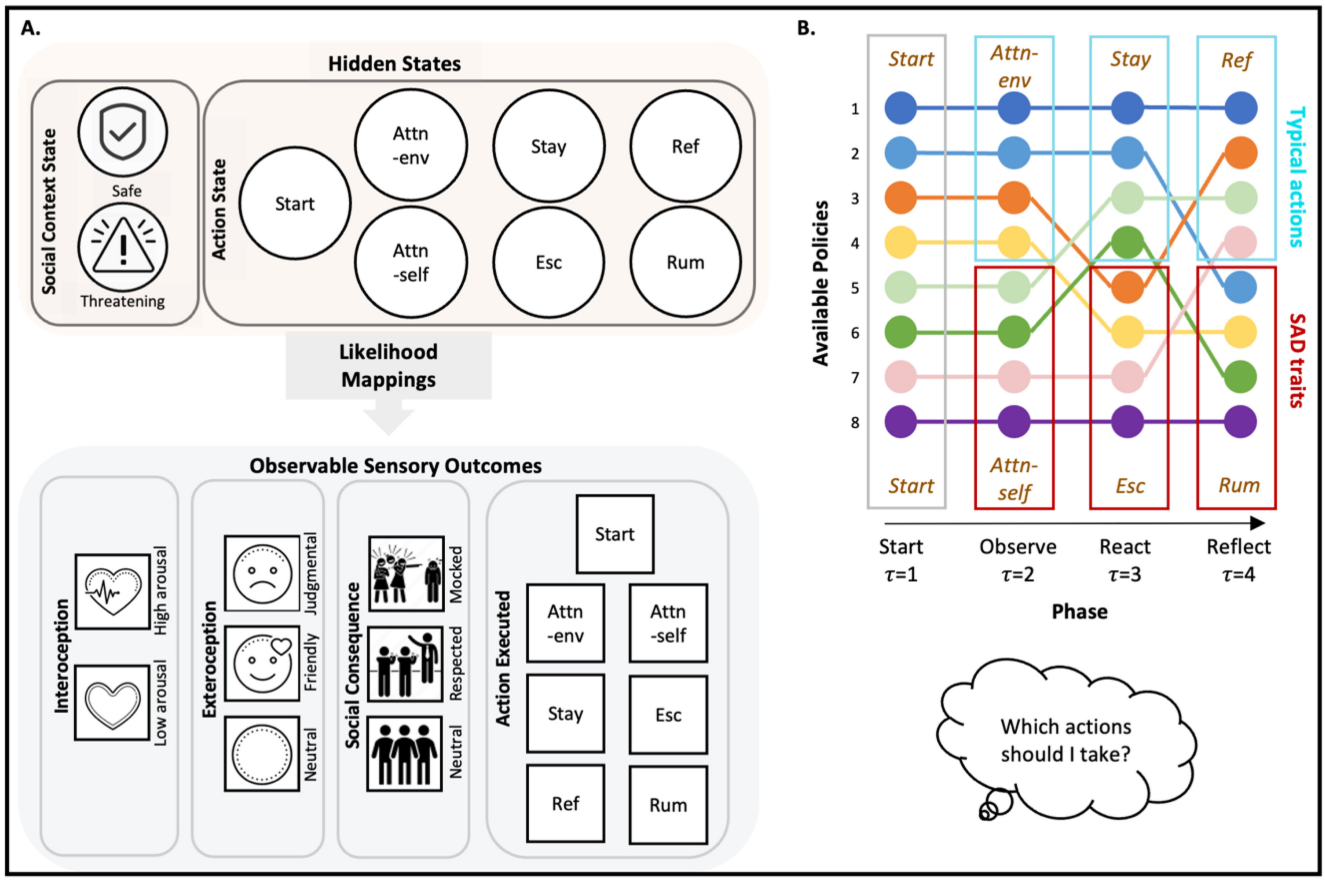
Multiple works have detailed the theoretical foundation and technical formulation of active inference (Friston et al., 2016; Sajid et al., 2021; Smith et al., 2022). Here, we provide a brief overview using language that is most relevant to the present study. Central to active inference is the generative model, in which statistical regularities between hidden state factors and observable sensory outcomes are stored and updated. The active inference agent[3] interacts with the environment through perception and action, which are tied to the statistical regularities encoded by the generative model.

Specifically, an active inference agent performs two key inferences. Firstly, they infer the hidden states that cause the observable sensory data (perception as inference) (Friston et al., 2015). Secondly, they infer the series of actions to be taken in order to achieve desirable sensory outcomes and avoid undesirable ones (action planning as inference) (Botvinick & Toussaint, 2012; Kaplan & Friston, 2018). These inferences are optimized by minimizing free energy, a quantity that measures the difference between the expected (or predicted) sensory outcomes and the observed ones. A low value of free energy indicates that the generative model can accurately predict the outcomes of the interaction between the self and the environment, thereby effectively guiding actions to bring about the expected observable sensory outcomes (Friston et al., 2016). Therefore, the process of minimizing free energy involves adjusting beliefs to better account for the observed sensory outcomes and selecting appropriate actions accordingly. Agents are also motivated to gain more information about the environment (Friston et al., 2015), as doing so will reduce free energy by improving the generative model's predictions. Technically, this involves evaluating different choices of actions or policies in terms of the expected free energy following that policy. The expected free energy subsumes expected information gain (a.k.a., epistemic affordance or value) and expected utility (a.k.a., instrumental affordance or value). In other words, following this free energy principle naturally motivates the agent to optimize both the epistemic value (through gaining information) and utility (through obtaining the desirable—i.e., unsurprising or preferred—sensory outcomes) (Sajid et al., 2021; Schwartenbeck et al., 2013). At any given moment, the agent's prior beliefs are shaped by past experiences and updated upon receiving new observable sensory outcomes (Friston et al., 2016). This ever-

[2] In this study, we use action (terminology commonly used in active inference) and behavior (typically used in CBT) interchangeably.

[3] Follow the common practice in using active inference, we use the word agent to refer to a simulated or real human being who takes an active role in producing their behaviors and perceptions.

Figure 1
Active Inference-Based CBT Model of SAD



Note. (A) A graphical depiction of the active inference model constructed based on the Hofmann (2007) conceptual model of SAD. There are two hidden state factors, namely the social context state (containing two possible states: safe and threatening) and the action state (containing seven possible states: start, attend to environment [Attn-env], attend to self [Attn-self], stay [Stay], escape [Esc], realistic reflection, and rumination [Rum]). There are four modalities of observable sensory outcomes, namely interoception (containing two possible outcomes: low and high arousal), exteroception (containing three possible outcomes: neutral, friendly, and judgmental expressions from the interlocutor), perceived social consequence (containing three possible outcomes: neutral, being respected, and being mocked), and executed action (containing seven possible outcomes which correspond to the seven action states). The probabilities of observing certain outcomes given certain hidden states are encoded in the likelihood mapping (see the online supplemental materials). (B) Eight possible policies (shown in different colors) resulting from seven available actions across the four timepoints. The active inference agent can pick a policy at Time Points 1–3 (there is no action decision to make at Time Point 4) to minimize their uncertainty about the environment while maximizing their desirable outcomes. When a policy is chosen, only the action immediately relevant will be executed. The top four actions circled in cyan (light gray) squares represent actions that are more adaptive, while the bottom four actions circled in coral (dark gray) squares represent actions commonly seen in SAD. CBT = cognitive–behavioral theory; SAD = social anxiety disorder; Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

evolving Bayesian process of inference, action, and belief updating is remarkably well-aligned with the CBT conceptualization of human behaviors.

From Conceptual to Quantitative Model of SAD

Below, we describe how the generative model of SAD was constructed (Figure 1). We first present the overall model structure. Concretely, the model construction was situated in a virtual environment used for the subsequent simulations. We then introduced how the key factors in the conceptual model were formalized in the quantitative model. The flexibility of active inference modeling allowed us to subsume or integrate these key factors of the conceptual model under a single goal; namely to minimize (expected) free energy, or equivalently, maximize model evidence. This is sometimes referred to as self-evidencing (Hohwy, 2016), where the active inference brain operates in a way that allows it to confirm their own predictions about their environment and themselves. For the concreteness of the model construction and simulation, we used exemplary descriptors when introducing the model components as detailed below. For example, we used "escape" to represent the general category of safety behaviors, and "stay (in the anxiety provoking setting)" to represent continued exposure. These two specific actions should be treated as examples of the large behavioral repertoire in real life. The same model structure was shared by all agents in all simulations. In the simulation, all agents experienced 100 social

encounters. For Simulations 1–7, the first 30 social encounters took place in a socially hostile environment, while the remaining 70 occurred in a socially supportive environment. For Simulation 8, the social environment is highly unstable or stochastic. In these simulations, agents were tasked to infer whether the social context was safe or threatening and take actions to optimize the outcomes of each social encounter.

Model Construction Rationale

Drawing from Hofmann (2007), we designed the simulation to incorporate a sequence of events that are typical in a social encounter. Each social encounter, considered as one trial, consisted of three phases: initialization, social interaction, and postevent processing. The social interaction phase was further divided into two time points: observation and reaction. Therefore, each social encounter consisted of four discrete time points (initialization, observation, reaction, and postevent processing).

In our model, an active inference agent needed to infer two hidden state factors in each trial—like humans, a simulated agent could not know the true state of the hidden world but could only infer it from observed sensory outcomes. The first hidden state factor was the social context state, indicating the probability of the current social context being safe (vs. threatening). The second hidden state factor was the action state, which encompassed several possible actions across different time points. At initialization, the action was simply to "start." At observation, agents orient attention either inwardly ("attend to self") or outwardly ("attend to environment"). At the time point of reaction, the options were to "stay" in the social setting or to engage in safety behaviors ("escape"). Finally, at the time point of postevent processing, the agent could either "realistically reflect" or "ruminate" about the social encounter. Combinations of these seven actions formed eight possible policies (Figure 1B). The term policy in the context of active inference means a sequence of actions that the agent can choose to take (Smith et al., 2022). At any given moment, multiple policies might be available to the agent, with each specifying a distinct sequence of actions. The agent assigns probabilities to available policies based on their knowledge about the environment, namely their predictions about outcomes that can be achieved by following each policy. A higher probability assigned to a policy indicates a better chance of achieving more desirable sensory outcomes by following that policy. At a given moment, an action is sampled from the marginal probability distribution over available policies (Smith et al., 2022). The evaluation and probability assignment processes for policies therefore become the basis of action inference.

The social context state and the action state jointly determined the probabilities of observing various sensory outcomes through probabilistic mappings (detailed below). Some outcomes (e.g., seeing a friendly expression and being respected) were desirable, while others (e.g., high arousal and being mocked) were undesirable. At each time point, the generative model yielded predictions about what sensory outcomes were likely to be observed given the current belief about the social context and action options. Thus, by inverting the generative model, an agent could infer the probability of the social context being safe through examining the observed sensory outcomes thus far. Similarly, action planning involved estimating the probability of observing different sensory outcomes in future time points. The planning was based on the current belief about the social context, one's own preferences (i.e., desirability of various outcomes), and the probability mapping between actions and observable sensory outcomes.

Vulnerability Factors of SAD

Next, we describe how major vulnerability factors of SAD were implemented in the current generative model. We identified and operationalized five vulnerability factors of SAD according to the Hofmann (2007) model (Table 1), namely (a) a preexisting bias of the social context being threatening, reflecting an initial level of social apprehension, (b) low self-efficacy in social settings, (c) altered social reward/loss processing, (d) heightened self-focused attention, and (e) postevent rumination.

First, an agent entered the initialization phase of the first social encounter with a preexisting belief about the probability of this social context being safe (vs. threatening). This prior probability was encoded in the parameter PriorSafe. A lower probability of the social context being safe indicated preexisting social apprehension, as seen in individuals with SAD (Clark & Wells, 1995; Gregory et al., 2016).

Second, during social interaction, a crucial factor influencing an agent's action was the probability of social mishaps. According to the conceptual model, the chance of social mishaps is linked to a person's self-image and social skills, which were together conceptualized as "self-efficacy," a widely studied factor in the literature (Iancu et al., 2015; Leary & Atherton, 1986). Low self-efficacy is related to anxiety and stress (Tahmassian & Jalali Moghadam, 2011) and is closely linked to more severe SAD symptoms (Iancu et al., 2015; Thomasson & Psouni, 2010). The level of self-efficacy was encoded by the parameter SEff, which influenced the probability of observing social mishaps, with social mishaps represented as observing the sensory outcome of "being mocked."

Anticipated social rewards versus loss can influence an individual's decision to stay or escape (Cremers, Veer, Spinhoven, Rombouts, & Roelofs, 2015). Catastrophizing social loss has been consistently linked to SAD (Cremers, Veer, Spinhoven, Rombouts, Yarkoni, et al., 2015; Rheingold et al., 2003; Richey et al., 2019). Furthermore, clinical examinations showed that individuals with SAD lacked motivation toward social rewards (Reilly et al., 2020). If the predicted reward value from a positive social outcome is substantially smaller compared to the potential social loss, individuals may choose avoidance behaviors (Hudd & Moscovitch, 2022; Richards et al., 2015). Thus, our third factor (a pair) was altered social reward and loss processing, encoded in SocGain and SocLoss respectively.

Next, we considered the two cognitive style factors. The fourth factor was attention allocation, encoded in SAttn, representing the tendency to engage in self-focused attention. Agents who tend to "attend to self" (a high positive value assigned to SAttn) would primarily observe and rely on interoceptive sensory outcomes (e.g., arousal level) rather than exteroceptive ones (e.g., facial expressions from interlocutors). Interoceptive information is less reliable than exteroceptive information in indicating the social context, potentially leading to misguided beliefs and actions (Schultz & Heimberg, 2008). Also, heightened self-focused attention could bias interoceptive perception toward negativity, for example, higher arousal (Bögels & Mansell, 2004; Mor & Winquist, 2002), leading to underestimation of one's social performance and overestimation of social threat (Gerrans & Murray, 2020; Mansell et al., 2003; Pineles & Mineka, 2005).

Table 1
Glossary of the Parameters Manipulated in the Simulations

Parameter name	Concepts in the conceptual model (Hofmann, 2007)	Interpretation in the computational model	Relevant active inference model component
PriorSafe	Social apprehension due to high perceived standards and poorly defined social goals	The agent's prior belief about the social context being safe when entering the simulation. Higher value indicates higher probability	$\boldsymbol{d}$
SEff	Perceived emotional control and perceived social skills Low self-efficacy leads to anticipating social mishaps	The agent's belief about having successful social experiences Higher value indicates higher confidence (and thus probability) of successful social encounters	$\boldsymbol{A}$
SAttn	Self-focused attention	The tendency of the agent to attend to themselves. Higher value indicates higher probability of observing the action of attending to themselves	$\boldsymbol{C}$
PostRum	Postevent rumination	The tendency of the agent to ruminate during postevent processing. Higher value indicates higher probability of observing the action of rumination	$\boldsymbol{C}$
Learning rate	Learning rate	Higher value indicates larger model update (relevant to $\boldsymbol{d}$ and $\boldsymbol{e}$) given certain observations	η
SocGain	Estimated social reward	The agent's preference of positive consequences during social encounters. Higher value indicates stronger preference	$\boldsymbol{C}$
SocLoss	Estimated social cost	The agent's aversion to negative consequences during social encounters. Larger negative value indicates higher aversion	$\boldsymbol{C}$

Lastly, during the postevent processing phase, choosing to "ruminate" (greater positive values of PostRum) would further consolidate or even exaggerate the negative experiences in social encounters (Mellings & Alden, 2000; Moberly & Watkins, 2008). Individuals who ruminate tend to possess prolonged negative thoughts and pessimistically predict future social performances (Brozovich & Heimberg, 2008), enhancing the chance of developing or exacerbating SAD symptoms (Kocovski & Rector, 2007; Rachman et al., 2000).

Constructing the Quantitative Model

After describing the building blocks of our generative model, we next specify the statistical relationships between these factors using the discrete space partially observable Markov decision process implementation of the active inference (Penny et al., 2006) (see Figure 1A for the graphic depiction, Tables 1 and 2 for the glossary of the model components, and online supplemental materials for the code and other technical details).

The two hidden state factors ($\boldsymbol{s}$) were:

1. Social context state with two possible contexts ("safe" and "threatening").
2. Action state with seven possible actions ("start," "attend to self," "attend to environment," "stay," "escape," (realistic) "reflect," "ruminate").

Observable sensory outcomes ($\boldsymbol{o}$) were generated from hidden states. The likelihood mappings between hidden states and observable sensory outcomes were encoded in the $\boldsymbol{A}$ matrices. In total, 15 possible observable sensory outcomes in four modalities were specified:

1. Modality 1 (interoception): using arousal level to represent interoception, which consisted of two possible outcomes ("high," "low").
2. Modality 2 (exteroception): using facial expressions from interlocutors to represent exteroception, which consisted of three possible outcomes ("judgmental," "friendly," "null").
3. Modality 3: representing the perceived social consequence with three possible outcomes ("mocked," "respected," "neutral").
4. Modality 4: allowing the agent to observe and keep track of actions executed, with seven possible observable sensory outcomes corresponding to the seven hidden action states ("start," "attend to self," "attend to environment," "stay," "escape," "reflect," "ruminate").

Correspondingly, the $\boldsymbol{A}$ matrices of different action states were specified as the following. The "start" state represented a baseline which, regardless of the social context state, generated the following outcomes: "low arousal," "null" facial expression, and "neutral" perceived social consequence.

We specified the effects of attention allocation as follows. Compared to under "attend to environment," under "attend to self," agents were more likely to observe high arousal, particularly

Table 2
Glossary of the Main Active Inference Model Components

Model component	General definition	Relevant model specification for simulations
$\boldsymbol{s}$	Hidden state factors	Two sets of hidden state factors: • s_1: action state • s_2: social context state
$\boldsymbol{o}$	Observable outcomes	Four sets of outcome modalities: • o_1: interoception • o_2: exteroception • o_3: perceived social consequence • o_4: observed action
$\boldsymbol{\pi}$	Distribution over policies	The probability distribution over eight allowable policies.
$\boldsymbol{D/d}$	$p(s)$. Prior probabilities over hidden states	Two D vectors for two hidden state factors: • D_1: prior probability distribution over action states • D_2: prior probability distribution over social context states the corresponding d vectors encode the corresponding beliefs in the agent's mind.
$\boldsymbol{A}$	$p(o\|s)$. Probability distribution that encodes the likelihood mappings between hidden states and possible outcome observations at each time point	Four $\boldsymbol{A}$ matrices for four outcome modalities: • A_1: likelihood mapping from hidden states to interoception • A_2: likelihood mapping from hidden states to exteroception • A_3: likelihood mapping from hidden states to perceived social consequence • A_4: likelihood mapping from hidden states to observed action
$\boldsymbol{B}$	$p(s_{\tau+1}\|s_\tau, \pi)$. Probability distribution that encodes the transition probability between hidden states	Two sets of $\boldsymbol{B}$ matrices for two hidden state factors: • B_1: transition probabilities for the action state • B_2: transition probabilities for the social context state
$\boldsymbol{C}$	$p(o_\tau)$. Probability distribution (via softmax conversion) that encodes the prior expectations over observable outcomes, essentially specifying the agent's preference or aversion for observing a given sensory observation	Four $\boldsymbol{C}$ matrices for four outcome modalities: • C_1: preference values for interoception • C_2: preference values for exteroception • C_3: preference values for perceived social consequence • C_4: preference values for observed action
$\boldsymbol{E/e}$	$p(p)$. Probability distribution that encodes the prior expectations over policies	A vector encoding prior expectations to choose from available policies

Note. Please refer to the online supplemental materials for more details on the formulation of the active inference model for the present simulations.

in threatening social contexts. Consequently, agents would be more likely to infer that the social context was threatening when observing high arousal under self-focused attention. Additionally, when agents attended to themselves, they would not observe expressions from interlocutors, which excluded their contribution to the inference of the social context. When agents attended to the environment, they were more likely to observe judgmental expressions in threatening contexts and friendly expressions in safe contexts. As a result, relying on exteroceptive outcomes, rather than interoceptive outcomes, would lead to an improved knowledge about the social environment.

As for behavioral reactions, we specified the likelihood mapping such that if agents chose to "escape," they would observe a "neutral" social consequence and "null" exteroceptive signals for both social context states, because they would no longer be in the social situation. Besides, "low arousal" was more likely to be observed when agents chose "escape." In contrast, if agents chose to stay, they would be more likely to observe themselves being "respected" with "low arousal" in a safe context and "mocked" with "high arousal" in a threatening context. In addition to the impacts from social context states, the likelihood of observing positive versus negative social consequences was further affected by self-efficacy, such that agents with lower self-efficacy were more likely to perceive themselves to be "mocked" than agents with higher self-efficacy (Iancu et al., 2015).

Finally, during postevent processing, when agents engaged in "rumination," they were more likely to observe outcomes that indicated a threatening social context regardless of what the true social context state was. In contrast, a realistic "reflection" would yield observable sensory outcomes that were reasonably indicative of the current social context state.

Throughout the social encounter, we allowed agents to observe their own actions. Accordingly, an identity matrix mapping from the hidden action states to the observed actions was implemented for agents to know which action was executed at each time point. The importance of this seemingly trivial point will be revealed next.

In active inference, there are no preferences for hidden states. Instead, agents can have preferences for various observable sensory outcomes generated from hidden states. These preferences are called prior preferences because they exist prior to the specific trial. One can treat these prior preferences like traits. Such preferences prescribe anticipation about the future, motivating agents to take actions that maximize the chance of being in states that are most likely to generate preferred outcomes (Pezzulo et al., 2018). Formally, these prior preferences were specified in the $\boldsymbol{C}$ vectors (e.g., values of the free parameters SocGain and SocLoss in the current study) and contributed to the probability over policies. Here, positive values were assigned to being "respected" as a successful social interaction was rewarding. Meanwhile, we assigned negative values to aversive outcomes including "high" arousal, observing "judgmental" expression from the interlocuter and being "mocked," which applied to all simulations, implying the agent's aversion toward these outcomes (see the supplemental code). Unless otherwise specified under a particular simulation, zero was assigned to the remaining outcomes, representing relatively neutral preference.

The prior preference coded in $\boldsymbol{C}$ does not merely imply reward or punishment. Rather, it reflects the extent to which agents expect to observe a specific outcome. A higher value for an outcome indicates that the agent has a higher expectation of observing that outcome. This reflects the individualistic functional values of observable sensory outcomes. For example, people who drink coffee more often than tea with breakfast would have a higher prior preference value for coffee than tea during breakfast, even if they generally prefer tea over coffee. Here, the two cognitive styles (i.e., self-focused attention and rumination) were translated into values in $\boldsymbol{C}$, with higher values indicating higher prior probabilities of observing themselves engaging in these thinking styles given their past experiences (Meehl, 1977; Shin et al., 2022).

In the following simulations, the prior belief about the social context was specific by the parameter (PriorSafe) in $\boldsymbol{d}$ vectors. Note that the true probability of the environmental state was encoded in the $\boldsymbol{D}$ vectors, which could be different from the agent's estimation. Thus, adaptive belief update over trials could be observed if the agent's belief about the social context encoded in $\boldsymbol{d}$ was updated to be more similar to values in $\boldsymbol{D}$.[4] In terms of the action state, the $\boldsymbol{d}$ vector simply specified that the agent always started with the "start" state.

The transitions of the hidden state factors were encoded in the $\boldsymbol{B}$ matrices. In the current study, we setup the transitions such that the social context state remained unchanged throughout a trial. In contrast, the agent was able to transit to different action state by selecting the corresponding action.

Prior probabilities over policies were specified in the $\boldsymbol{e}$ vector, indicating the agent's prior preferences over policies—reflecting action habit—in the beginning of each trial. Over trials, the agent could update their preference over policies. Habit could thus emerge from the increased prior preferences for certain policies.

Finally, the speed of the updates in belief $\boldsymbol{d}$ and habit $\boldsymbol{e}$ was influenced by the active inference learning rate parameter η, which was set to 1 by default and kept stable throughout the 100 trials for all agents.

We stress again that the specific descriptors used to label the hidden states and outcomes in the present study should be treated as examples to construct a concrete model. For example, the "escape" behavior representing safety behavior can be replaced by other safety behaviors, for instance, eye contact avoidance. The same logic applies to any other concrete descriptors in the present model.

We further elaborated on how the generative model functioned throughout the 100 trials in the first simulation below.

Simulations

Eight simulations were conducted. We first demonstrated the beliefs and behaviors of a healthy agent, and then investigated how the five vulnerability factors listed above led to biased beliefs about the social context and maladaptive behaviors. We manipulated one parameter at a time while keeping other parameters the same as in the healthy agent to test the specific effects of each vulnerability factor in isolation (Tables 1 and 3).

All simulations started with a flat $\boldsymbol{e}$, indicating no preexisting habit. See Tables 1 and 3 for a more detailed record of parameter settings. Additionally, two supplementary simulations, exploring the influence of different learning rates η and the influence of a social environment that changed twice (from supportive to hostile to supportive), were included in the online supplemental materials. Note that the parameter settings for each simulation represent the initial beliefs and preferences of our simulated agents, that is they are the prior beliefs entailed by the agents when entering the simulated context. This speaks to a key point that every simulated agent is unique because they have their own set of initial beliefs about the world and personal preferences before entering the simulated social context. These individual differences are captured in their unique generative models through different parameter settings. As we will see in the discussion, it is possible to recover these parameters that characterize any given person, using Bayesian model inversion.

Each simulated agent completed 100 trials of social encounters. In Simulations 1–7, for the first 30 trials, a hostile social environment was simulated (i.e., with a probability ratio of .1: .9 for the social context of a given trial being safe: threatening); for the 31st to the 100th trial, this probability ratio was reversed (safe: threatening = .9: .1). In the eighth simulation, the social environment was set to be stochastic throughout the 100 trials (with the probability ratio for safe: threatening = .5: .5).

Results

Active Inference-Based CBT Model of SAD

In this section, we describe the resultant active inference account of the Hofmann (2007) model of SAD. The core of the active inference framework is the generative model in a person's mind. Adaptive or maladaptive behaviors are products of this generative model. We constructed a generative model incorporating all major factors in Hofmann's (2007) conceptual model (Figure 1). This means that by inputting specific values (Table 3), a simulated participant (referred to as an agent) equipped with this generative model can exhibit behaviors similar to individuals experiencing SAD. This generative model itself is not explicitly or consciously available to an agent. One can regard this generative model as an implicit belief system that includes thoughts and yields actions (Smith et al., 2021).

[4] In the SPM active inference scheme, small letters denote Dirichlet concentration parameters and the Dirichlet distribution is the conjugate prior of the categorical distribution (see Smith et al., 2022 for technical details).

Table 3
Parameter Values Used in the Simulations 1–7

Simulation	1. Healthy agent	2. Prior bias	3. Low self-efficacy	4. Social reward/loss dysfunction	5. Heightened self-focused attention	6. Rumination	7. Multiple vulnerabilities to SAD
PriorSafe	.6	**.1, .3, .5**	.6	.6	.6	.6	**.2**
SEff	.8	.8	**.2, .4, .6**	.8	.8	.8	**.2**
SAttn	0	0	0	0	**1, 2, 3**	0	**3**
PostRum	0	0	0	0	0	**1, 2, 3**	**3**
SocGain	1.1	1.1	1.1	**0.9, 0.7, 0.5**	1.1	1.1	**0.5**
SocLoss	−1.3	−1.3	−1.3	**−2, −2.5, −3**	−1.3	−1.3	**−3**

Note. The parameters in Simulations 2–7 that differ from those in Simulation 1 are given in bold text. SAD = social anxiety disorder. Please refer to Table 1 for detailed definition of the model parameters.

Our simulation consisted of 100 social encounters such that each social encounter was one trial consisting of four discrete time points, (t1) initialization, social interaction, including (t2) observation and (t3) reaction, and finally (t4) postevent processing. In each social encounter, an agent needs to infer whether the social setting is safe or threatening, based on observable outcomes, and execute actions accordingly. The first state factor is the social context state, which encompasses two possible states (i.e., safe and threatening) across all time points. For example, in the timeline above, observing a judgmental expression (t2) and being mocked (t3) may lead to the inference that the social context that yields these observations is more likely to be threatening than safe. The second state factor is the action state, which encompasses several possible actions across different time points. At initialization (t1), the action is simply to "start." At observation (t2), the agent can either (a) "attend to self" or (b) "attend to environment." At reaction (t3), the options are to (a) remain in the social setting ("stay") or to (b) engage in safety behaviors ("escape"). Finally, at postevent processing (t4), the agent can either (a) (realistically) "reflect" or (b) "ruminate" about the social encounter. Combinations of these seven actions form eight possible policies (i.e., a fixed sequence of actions; see Methodology section; Figure 1B). Importantly, as shown in Figure 1A, different combinations of the inferred social context state and the action state would yield different predicted observable sensory outcomes. For example, given a threatening social context, an agent may predict outcomes like observing high arousal when "attend(ing) to self" and observing judgmental expressions when "attend(ing) to the environment" during observation. Further, staying in the social setting would lead to negative social outcomes (represented as "being mocked"). Each of these observable outcomes has a value, with positive values indicating pleasant/desirable experiences and negative values indicating unpleasant/unwanted experiences. At a given moment, an agent can infer which social context state they are likely in based on which social context state better explains their observed sensory outcomes. In the example above, if the agent observes judgmental expressions, then they are more likely to infer that the social context is threatening. The agent then plans what actions to take to obtain desirable future sensory outcomes (action planning as inference). These predicted outcomes may or may not match with the actual outcomes (see Methodology section for the state-outcome mapping), because the actual outcomes depended on the interactions between the true social context state, which is importantly hidden from the agent, and the executed actions. Inherently, an active inference agent is motivated to improve model predictions by gathering more information about the environment while trying to obtain better outcomes (see Methodology section and the online supplemental materials for detailed discussion).

Within this model structure, we quantitatively operationalized the five key factors of SAD according to Hofmann's (2007) model (see the Methodology section and Tables 1 and 3). First, an agent enters the social environment with a preexisting belief about the probability of this social context being safe (vs. threatening), encoded in the parameter PriorSafe. A lower value indicates preexisting social apprehension. Second, during social interaction, the level of self-efficacy is encoded by the parameter SEff, which influences the probability of observing social mishaps (represented as observing the sensory outcome of "being mocked"). The third factor (a pair) is altered social reward and loss processing, encoded in SocGain and SocLoss, respectively. Smaller SocGain indicates lower value gains from positive social outcomes, and larger SocLoss indicates magnified costs of negative social outcomes. The fourth factor is attention allocation, encoded in SAttn (higher value indicating higher tendency to attend to the self). This factor affects whether the agent will observe interoceptive versus exteroceptive sensory outcomes. When "attend(ing) to self" the agent will primarily observe and rely on interoceptive sensory outcomes (e.g., arousal level) rather than exteroceptive ones (e.g., facial expressions from interlocutors). Lastly, during the postevent processing phase, choosing to "ruminate" rather than realistically "reflect" on the social encounter (greater positive values of PostRum) will lead to higher probability of observing negative experiences.

Different Social Belief and Behavioral Trajectories Under the Influences of Different Factors

With this generative model, we systematically investigated the effects of different factors on the agent's (a) estimated probability of the social context being safe in the beginning of each social encounter (here onwards referred to as social belief) and (b) avoidance (escape) versus approach (stay) behaviors. We chose these two markers, one representing social apprehension cognition and the other representing avoidance behavior, as they are important manifestations of SAD. In the first seven simulations, agents completed 100 trials of social encounters with the same social environment setting. Specifically, in the first 30 trials, a hostile social environment was simulated (i.e., with a probability ratio of .1: .9 for the social context of a given trial being safe: threatening). From the 31st to the 100th trial, this probability ratio was reversed: p(safe): p(threatening) = .9: .1. This change from a hostile to a supportive environment required the agent's adaptability. In the eighth

simulation, we employed a stochastic social environment: p(safe): p(threatening) = .5: .5.

Simulation 1: The Healthy Agent

In this baseline simulation, we demonstrated how an active inference agent without vulnerability to SAD—thus deemed healthy—would behave throughout the 100 social encounters (trials). Specifically, this healthy agent entered the first trial with a small positively biased belief about the social context (PriorSafe = .6, indicating the prior of p_{safe}: $p_{\text{threatening}}$ = .6: .4 in the agent's mind), reflecting the optimistic bias in people without psychological disorders (Sharot, 2011; Weinstein, 1989). This prior probability about the social context at the beginning of each trial could be updated over trials. The agent was also equipped with a high self-efficacy level (SEff = .8). Furthermore, this agent assigned moderate utility values for negative and positive social consequences (SocLoss = −1.3, SocGain = 1.1), with the negative one having slightly larger magnitude to reflect the typical effect that loss looms larger than gain. Finally, this healthy agent had no prior preference for or against self-focused attention (SAttn = 0) or rumination (PostRum = 0) (Tables 1 and 3). Please see online supplemental materials for a step-by-step description of the healthy agent' simulation process.

As expected, in the initial 30 trials (i.e., the hostile environment with high probability of being threatening), the healthy agent quickly increased their tendency of escaping (Figure 2A, left), leading to frequent neutral outcomes in terms of social consequences. Occasionally, when the agent stayed in the social setting, they observed being mocked. In the subsequent 70 trials where the environment changed to be supportive, with a high probability of being in a safe social context on a given trial, the agent was able to change the action to stay in the social setting, demonstrating adaptability. The agent thus gained frequent social rewards from successful interactions. It is also noteworthy that occasional negative outcomes ("being mocked") during the latter 70 trials were often followed by escaping in the subsequent trial, but this tendency to escape did not persist. In terms of social belief, the agent successfully updated their belief about the social context by adopting a suitable strategy that involved attending to the external environment, adequately exposing themselves to social situations, and realistically reflecting on experiences, the agent successfully updated their belief about the social context (Figure 2B, left). The agent was also able to optimize policy selection toward the end of the simulation (Figure 2C, left).

Notably, optimal information processing strategies ("attend to environment" and realistically "reflect" on the social encounter experience) led to accurate inferences and adaptive actions in social situations. It is also evident that both reducing uncertainty (gaining knowledge about the environment) and increasing utility (maximizing the chance of social reward and minimizing the chance of social loss) were crucial during action planning. To develop an accurate understanding of the social context while optimizing the outcome, the agent needed sensory evidence from the most informative source (i.e., the environment) and sufficient exposure to the social settings.

Simulation 2: Effects of a Preexisting Negative Belief About Social Environment

In this simulation, we investigated how negatively biased prior beliefs about the social context (Clark & Wells, 1995; Hofmann, 2007) would impact agent's belief update about the social context and tendency of safety behaviors. A daily life example would be an adolescent who recently transferred to a new school was warned by their parents that their new peers may be judgmental. This adolescent thus enters the first day of school with a negatively biased prior belief while carrying no other risk factors. Here, instead of assigning an optimistic bias to the healthy agent as in Simulation 1 (Sharot, 2011), we simulated two agents with negatively biased prior beliefs and one agent with an unbiased prior belief (PriorSafe = .1, .3, .5; Tables 1 and 3). All other settings were kept the same as for the healthy agent.

We hypothesized that compared to the healthy agent, agents with lower PriorSafe would show more exaggerated negative beliefs about the social context at the end of the first 30 trials and choose "escape" over "stay" more often. Their belief update in the supportive environment would then be slower. However, given that the other factors were "intact" (i.e., the same as a healthy agent), these agents would still be able to eventually show adaptive actions and gradually correct their belief about the social context.

The results showed that when the true social environment was hostile (the first 30 trials), all three agents adaptively escaped from the social contexts frequently. When the social environment changed to be supportive, these agents were able to adapt their actions and chose to stay in the social context more often. However, compared to the healthy agent, they still showed more safety behaviors, especially the one with the most negative prior belief (Figure 3A). In terms of information sampling, like the healthy agent, these agents attended to the environment instead of themselves most of the time. They also engaged in realistic reflection rather than rumination during postevent processing. Such an action profile would allow them to obtain information about the social context and inform their belief updating. Indeed, their beliefs were updated over the 100 trials. In the first 30 trials, an initial negatively biased prior belief was well matched to the true hostile social environment and was further strengthened. When the environment was hostile, compared to agents with less extreme negative prior beliefs, there was an obvious overshoot toward negativity for the agent with the most extreme negative prior belief (PriorSafe = .1). This was followed by a slower update toward positivity after the true social environment changed to be supportive.

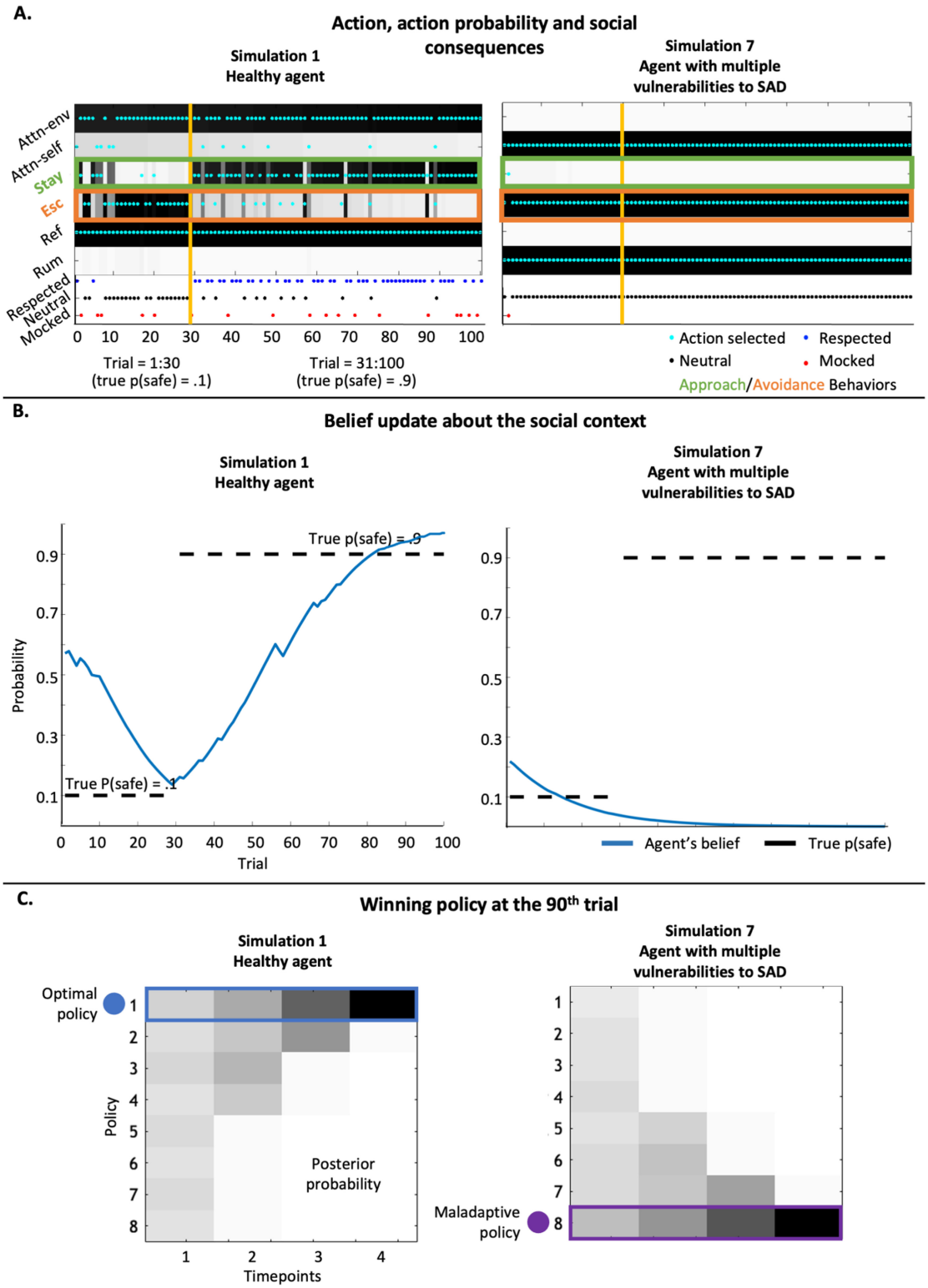
Overall, results in this simulation suggested that keeping everything else equal to the healthy agent, agents having negatively biased beliefs about the social context would be able to gradually update their belief and reduce safety behaviors in a supportive environment, but they may require more corrective experiences.

Simulation 3: Effects of Self-Efficacy

In our quantitative model, agents with lower self-efficacy would be more likely to observe negative social outcomes. Lower social self-efficacy may be due to social skill deficits. Here, we emphasize the self-perception of lower efficacy (e.g., people consider themselves to be "socially awkward or incompetent"). In this simulation, we investigated the impact of self-efficacy on belief update and action selection. We simulated agents with various levels of self-efficacy (SEff = .2, .4, and .6 from low to high) while other parameters were kept the same as for the healthy agent. We predicted that

Figure 2

Simulation Results for the Healthy Agent and the Agent With Multiple Vulnerabilities to SAD



Note. (A) The selected action and perceived social consequences across the 100 trials for the healthy agent (left) and the agent with multiple vulnerabilities to SAD (right). The yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all

(*Continued on next page*)

with lower self-efficacy, the agent would be more likely to predict and experience social mishaps when staying in the social setting, particularly when the social environment was truly hostile. Thus, agents with lower self-efficacy would be more likely to choose "escape" in the social encounter (Heimberg et al., 2010), even more so in the hostile environment than in the supportive environment. We also predicted that agents with lower self-efficacy would be able to update their belief about the social context in the supportive environment, but with a rate much slower than agents with higher self-efficacy.

As shown in Figure 4A, lower self-efficacy was indeed associated with a lower probability and frequency of choosing to stay in the social setting. At a considerably low self-efficacy level (SEff = .2), the agent exhibited a strong tendency for avoidant action, regardless of whether the social context was safe or threatening. At a moderate level (SEff = .4), the probability of choosing to stay increased, although the agent's behavior was still dominated by escaping. For these first two agents, while they successfully avoided social loss in a hostile environment, they almost entirely missed social gains in the supportive environment. In contrast, the agent was much more likely to stay in the social setting when their self-efficacy was higher (SEff = .6) and thus obtained more social gains. When considering belief updating (Figure 4B), all three agents remained attentive to the environment in most trials. As a result, they were able to gather exteroceptive information about the social context and adjust their beliefs accordingly. Agents with low self-efficacy quickly learned that the social context was threatening but were able to detect changes in the environment based on exteroceptive information, such as a smile from the interlocutor, in the supportive environment. While choosing to stay could allow them to gather further empirical evidence about social consequences, choosing to escape forfeited this opportunity. Consequently, the belief updating toward positivity was slower for agents with lower self-efficacy compared to those with higher self-efficacy. This indicated that excessive escaping prevented the agent from gaining information based on the social consequences. The divergence between successful social belief update and persistent escaping action indicated that individuals with sufficiently low self-efficacy, lacking confidence in acquiring positive social outcomes, would prefer the safe option of escaping, even when they know that the environment could be accommodating.

This simulation demonstrated that low self-efficacy alone was sufficient to generate and sustain avoidance behaviors, and to a milder degree exaggerate negative beliefs about the social context.

Simulation 4: Effects of Attenuated Social Reward and Catastrophized Social Loss

In this simulation, we investigated the impact of impaired sensitivity to social reward and loss on belief update and action selection. To investigate the consequences of altered appreciation for positive and negative social consequences, we conducted simulations using agents that assigned relatively small positive values of 0.5, 0.7, 0.9 and large negative values of −3, −2.5, −2 to SocGain and SocLoss, respectively. We predicted that agents having larger negative values for social loss relative to social gain would be prone to develop persistent safety behaviors.

As the magnitude of social loss increased relative to social reward, agents in this simulation consistently exhibited excess safety behaviors across the 100 trials (Figure 5). The probability of choosing to escape from social interactions increased as social gain decreased and social loss increased. However, these agents' beliefs about the social context were not particularly biased toward negativity. Despite the lack of a strong belief distortion, these agents struggled to overcome their tendency to escape due to their prediction of catastrophic social loss and quickly developed the habit of escaping. These catastrophic predictions and habitual avoidant actions hindered the agent's ability to break from excessive safety behaviors and adapt to the environment. Meanwhile, as the predicted gain from successful social interactions was relatively lower than the loss of negative social outcomes, the agent was not motivated to change their behavioral patterns.

This simulation further highlighted the potential mismatch between social beliefs and action tendency. When an individual has a high aversion to social loss relative to potential gain, they are likely to engage in avoidance behaviors, even if their inferred safety level of the social context is accurate.

Simulation 5: Effects of Heightened Self-Focused Attention

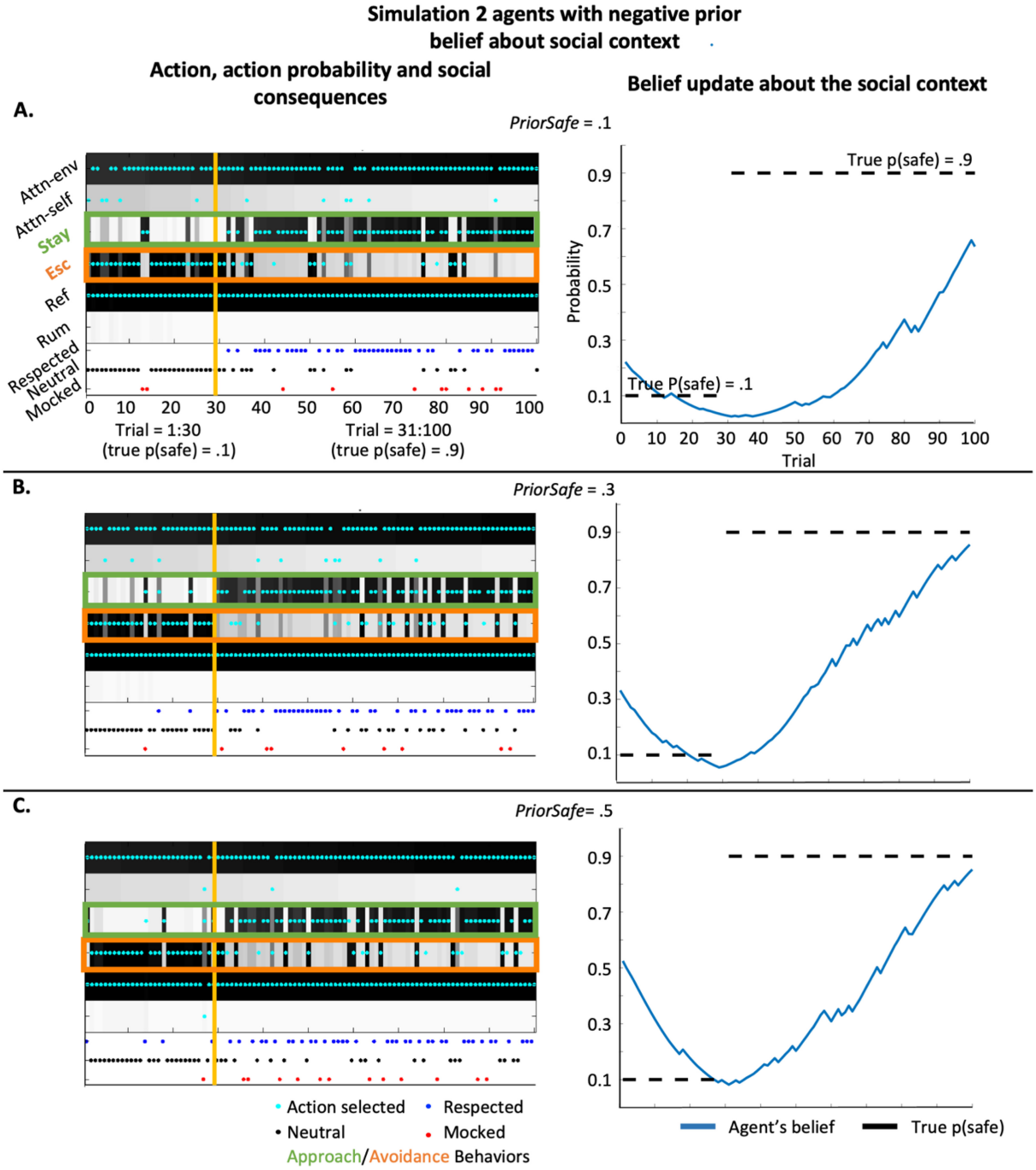
Self-focused attention has long been a crucial factor in CBT models of SAD (Spurr & Stopa, 2002), preventing people from effectively receiving information from the external social environment. In this simulation, we investigated the impact of heightened self-focused attention on belief update and action. For agents with extremely heightened self-focused attention, SAttn was set to 3. This indicated their high expectation of observing inwardly focused attention We also simulated two other agents with less extreme values (SAttn = 2 and 1). We predicted that agents with heightened self-focused attention would be vulnerable to maladaptive beliefs

Figure 2 (*continued*)

panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. (B) Belief update about the social context. Blue (gray) lines depict how the agent's belief about the social context evolved throughout 100 trials. Black dotted lines mark the true probability of the social context being safe. (C) The policy update and preference at the 90th trial. At each timepoint of each trial, the posterior probability of the most preferred policy is updated. Here, the healthy agent learned the optimal policy (Policy 1) while the agent with high vulnerability to SAD preferred the maladaptive one (Policy 8). Please also see Figure 1 for details. SAD = social anxiety disorder; Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

Figure 3

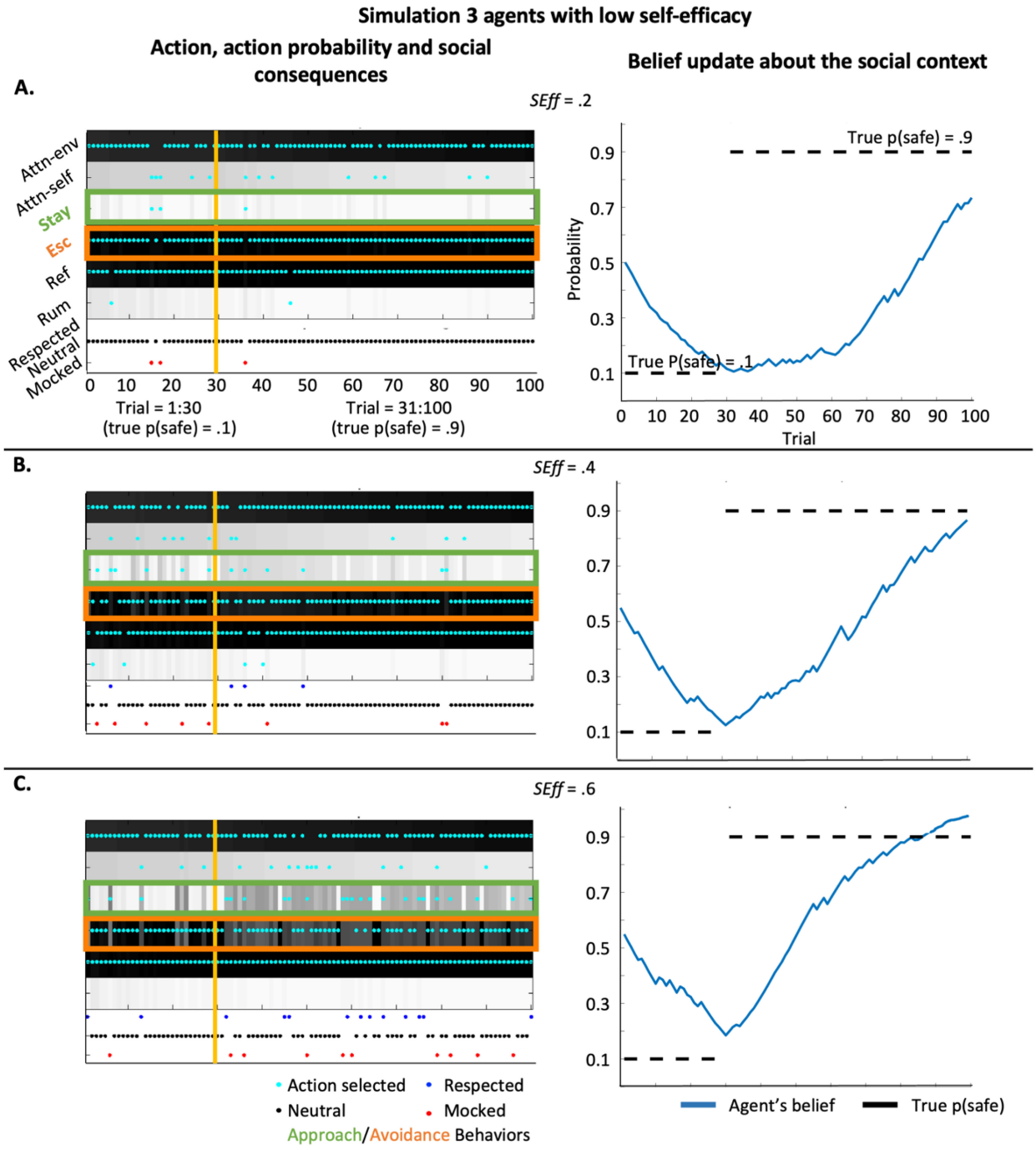
Action Selection, Action Probability, Perceived Social Consequences, and Belief Update for Simulation 2



Note. Results for agents with different prior beliefs about the environment. (A) PriorSafe = .1, (B) PriorSafe = .3, (C) PriorSafe = .5. Left panels: Action probability, action selection, and perceived social consequences. As in Figure 2, the yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4 and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

Figure 4

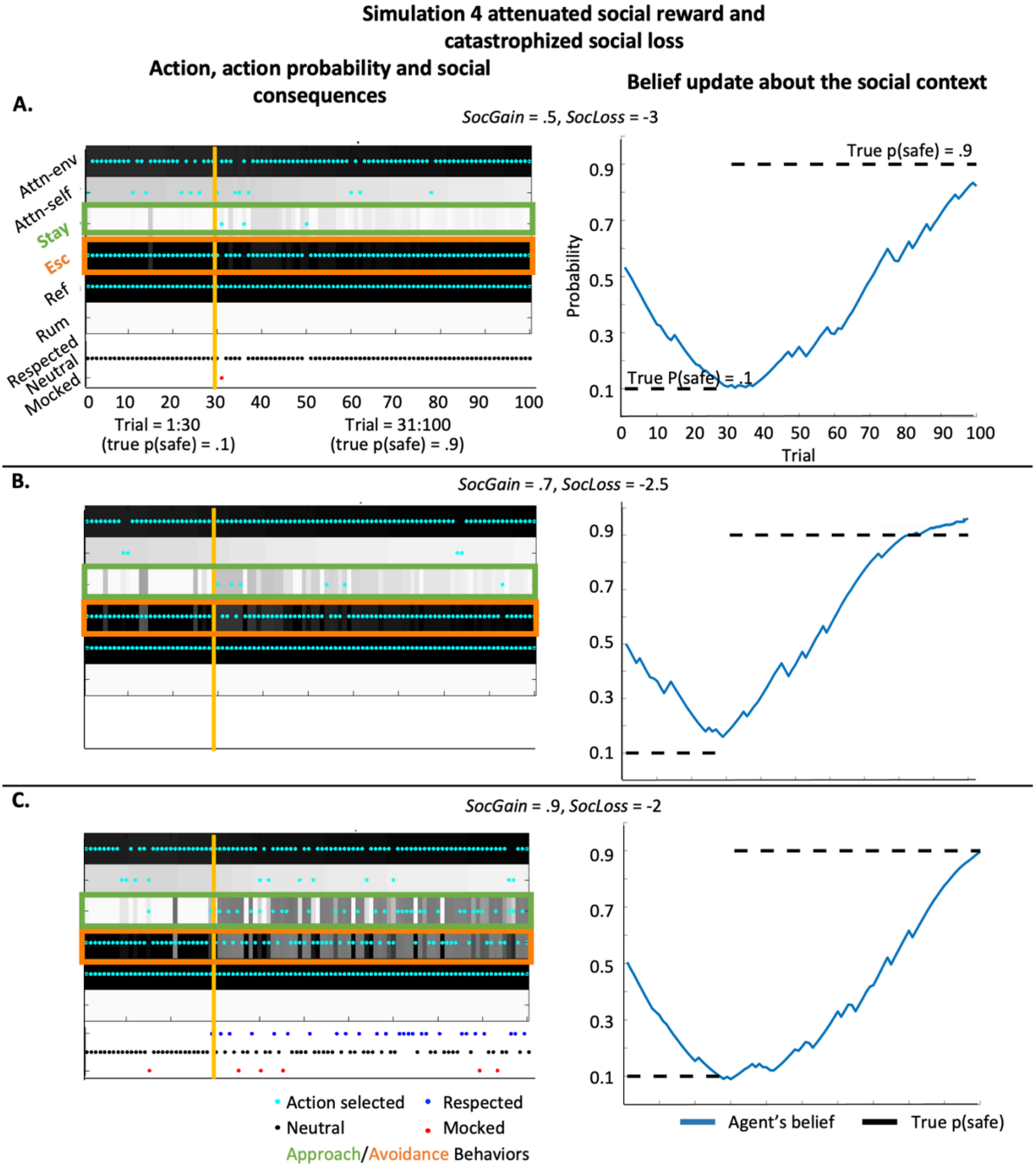
Action Selection, Action Probability, Perceived Social Consequences and Belief Update for Simulation 3



Note. Results for agents with different levels of low self-efficacy. (A) SEff = .2, (B) SEff = .4, (C) SEff = .6. Left panels: Action probability, action selection, and perceived social consequences. As in Figure 2, the yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4, and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

Figure 5

Action Selection, Action Probability, Perceived Social Consequences, and Belief Update for Simulation 4



Note. Results for agents with different perceived social reward and loss. (A) SocGain = 0.5, SocLoss = −3, (B) SocGain = 0.7, SocLoss = −2.5, (C) SocGain = 0.9, SocLoss = −2. Left panels: Action, action probability, and perceived social consequences for agents with different levels of abnormal reward processing, with the agent with the most severe reward dysfunction shown in (A). As in Figure 2, The yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4, and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

given their suboptimal information source, leading to inaccurate inferences about the social context and less optimal actions.

Figure 6 illustrates that agents with heightened self-focused attention largely stayed in the threatening context and occasionally escaped from the safe context, resulting in an action pattern that was not beneficial considering the potential social consequences (marked by the red dots in the first 30 trials and black dots in the last 70 trials in the figure). With heightened self-focused attention, their decision to stay or escape was more stochastic, indicating that their actions were not well informed by the predicted social context. This maladaptive behavioral pattern could be explained by their reliance on interoceptive rather than exteroceptive sensory outcomes, as the former provides less accurate information about the social context. Consequently, they experienced frequent social mishaps in the first 30 trials of the simulation. Also due to this lack of accurate information, the agent's belief about the social context was unable to align with the true context states throughout the simulation (Figure 6, right).

Overall, agents with heightened self-focused attention had difficulty inferring the accurate social context state. This lack of information hindered their action decisions, yielding undesirable consequences throughout the simulation. This simulation highlighted the importance of employing cognitive strategies that facilitate accurate information processing.

Simulation 6: Effects of Rumination

In this simulation, we investigated the impact of rumination on belief update and action. Like self-focused attention, the tendency to ruminate was encoded in a parameter PostRum in $\boldsymbol{C}$. Here, PostRum was set to 3, 2, and 1, reflecting different severity levels of rumination. The higher value indicates higher rumination tendency. We predicted that agents with higher rumination tendency would be more likely to stick to safety behaviors after occasional social mishaps, as their perceptions about the environment would be negatively biased by the rumination process.

As shown in Figure 7, when the rumination tendency was low (PostRum = 1; Figure 7, Panel C), rumination had relatively smaller impacts compared to other factors. When rumination tendency was higher (Panels A and B) the probability of escaping even in a supportive environment was increased. Furthermore, agents with a very high tendency to ruminate exhibited a slower belief update toward positivity in the latter 70 trials (Figure 7, right). This resulted in a lower confidence in the social environment of being safe throughout the last 70 trials. With such negative beliefs about the social environment, the agent with higher rumination tendency reasonably preferred to escape even when the environment was supportive. Overall, this simulation demonstrated that agents with high levels of rumination were more prone to safety behaviors in supportive social environments and were more vulnerable to developing symptoms after occasional social mishaps.

Simulation 7: Cumulated Effects of Multiple Vulnerability Factors

In our final simulation within the first type of social environment setting, we introduced an agent possessing multiple vulnerability factors. This prototypical agent possessed a negative prior belief about the social context (PriorSafe = .2) and low self-efficacy (SEff = .2) in the beginning of the simulation. Furthermore, this agent was less sensitive to social rewards (SocGain = 0.5) and much more sensitive to social punishments (SocLoss = −3). Additionally, this agent preferred cognitive styles of self-focused attention and rumination (SAttn = 3, PostRum = 3). This agent was hypothesized to suffer from extremely negative social beliefs and safety behavior patterns that were resistant to change even after the environment turned to be supportive.

As depicted in Figure 2, this agent who scored high on all five cognitive vulnerability factors was characterized by a fixed and exaggerated negative social belief and persistent safety behaviors. Also, in marked contrast to the healthy agent, the agent with multiple vulnerability factors quickly developed a preference for a maladaptive policy involving avoidance, promoted self-focused attention and rumination. Furthermore, they sticked to this policy throughout the 100 trials and showed an inability to adaptively adjust their action patterns. Without sufficient information from the external world to correct the negative beliefs, the agent persisted with a precise belief that they were in a threatening social context and chose to escape across trials. This simulation demonstrated how maladaptive beliefs and behaviors in SAD are developed and maintained in individuals with high vulnerability.

Simulation 8: An Alternative, Stochastic Social Environment

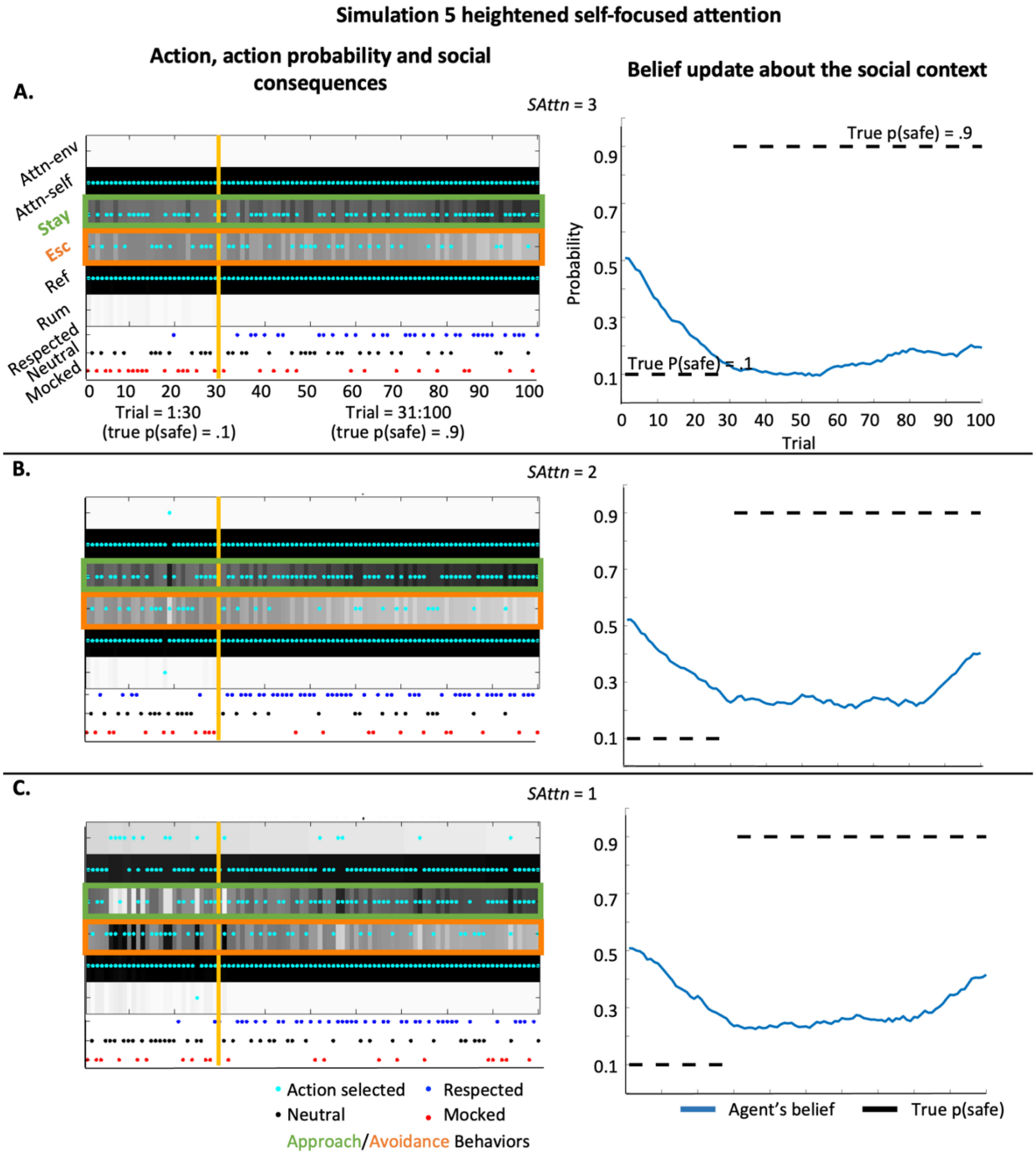
To demonstrate the applicability of our modeling approach in different social environments, we simulated the cognitive–behavioral performance of a healthy agent and an agent with low self-efficacy (SEff = .2) in an alternative, stochastic environment. Each simulated agent completed 100 trials of social encounters in an environment that was equally likely to be safe or dangerous in each trial—that is, the trial-level probability ratio of p(safe): p(threatening) = .5: .5. The model parameter values for the two simulated agents were kept the same as in Simulation 1 for the healthy agent, and Simulation 3 for the agent with low self-efficacy.

Similar to Simulation 3, the agent with low self-efficacy exhibited excessive escaping behaviors, resulting in limited social exposure regardless of the environment context (Figure 8, Panel B left). The safety behaviors stemmed from their perceived inadequacy in social competency. Consequently, the low self-efficacy agent consistently avoided social interactions, failing to acquire sufficient information to update their beliefs about the current social context. Despite entering the first trial with slightly positive initial beliefs, such an agent tended to exaggerate the negativity of the social context in a highly stochastic environment over the course of trials (Figure 8, Panel B right). This progressive negativity in beliefs further exacerbated the excessive avoidance in low self-efficacy agents, perpetuating SAD-like cognitive and behavioral patterns. In contrast, the healthy agent entered the first trial with a slightly optimistic expectation about the social context, and the initial rewarding social outcomes boosted this agent's belief about the social context, making it more positively biased. However, the agent was able to update their belief and eventually reach a relatively accurate judgment about the social context. This belief update allowed the healthy agent to make informed decisions regarding their actions, leading to an overall balanced decision pattern between staying and escaping.

This simulation further highlights the maladaptive nature of low self-efficacy in social anxiety, demonstrating how it can cause consistent impairments in an individual's ability to engage in social

Figure 6

Action Selection, Action Probability, Perceived Social Consequences, and Belief Update for Simulation 5



Note. Results for agents who differed in self-focused attention. (A) SAttn = 3; (B) SAttn = 2; (C) SAttn = 1. Left panels: Action, action probability, and perceived social consequences for agents with different levels of heightened self-focused attention. As in Figure 2, the yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4, and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

interactions and gain social rewards regardless of the social environments.

Discussion

Constructing quantitative formulations of psychopathology is pivotal for advancing clinical science (Meehl, 1977; Montague et al., 2012). Recognizing the theoretical alignment between CBT and active inference, we took the first step to construct a quantitative CBT model of psychopathology, using SAD as an example. We first translated Hofmann (2007), a well-established CBT conceptual model of SAD, into a quantitative mechanistic model, by clarifying psychological constructs and operationalizing them within the active inference framework. Through a series of simulations, we uncovered how negatively biased beliefs regarding social contexts and avoidance behaviors can develop as results of Bayesian belief updating. Most importantly, the active inference-based CBT model allowed us to accommodate a wide range of interconnected mechanistic factors of SAD within the same framework while quantitatively evaluating their effects on beliefs about social context and avoidance behaviors. The model-building process requires rigorous delineation and statistical representation of conceptual factors. This intrinsic feature of the model-building process facilitates clarification of several previously ambiguous psychopathological pathways toward SAD. Below we first discuss how these findings enhance our understanding of SAD psychopathology and hold the potential to facilitate client-centered intervention identification, followed by an elaboration on broader implications and applicability of employing active inference to formalize other clinical theories.

Differential Influences of Vulnerability Factors and Implications for Personalized Intervention

Computational modeling of the mechanisms enables thought experiments by systematically manipulating specific factors, which can be challenging to achieve when relying on real human data (Gratch & Marsella, 2004). It provides a unique opportunity to disentangle the complex interactions between various factors in a conceptual model, quantitatively delineating the unique impacts of each factor and thereby informing intervention. In this study, we tested five key factors proposed in the conceptual model of SAD (Hofmann, 2007). Our findings highlight how certain factors can directly lead to avoidance behavior, while others having a larger impact on social belief.

Among these factors, low self-efficacy and larger perceived social loss relative to social gain were shown to directly induce safety behaviors. Self-efficacy is closely associated with SAD in both children and adults (Bandura et al., 2003; Rudy et al., 2012). Socially anxious individuals often show lower confidence in their social performance (Harrison et al., 1997). Accordingly, we formalized self-efficacy as the likelihood of observing negative social consequences. Simulation results quantitatively confirmed the significant role of low self-efficacy in generating avoidant behaviors and slowing down adaptive belief updating in supportive, hostile, as well as stochastic environments. Crucially, agents with low self-efficacy exhibited excessive avoidant behaviors even without negatively biased beliefs about the social context. This implies that individuals with very low self-efficacy may anticipate negative social consequences and opt for safety behaviors even when they know the context is not threatening. This result provided a potential explanation of how targeting self-efficacy may uniquely contribute to better treatment outcomes in SAD (Gaudiano & Herbert, 2003). Values of positive and negative social consequences are another pair of factors crucial in SAD psychopathology (Caouette & Guyer, 2014). Individuals with SAD exhibit reduced sensitivity to social gains with diminished motivation to approach potentially rewarding stimuli (Cremers, Veer, Spinhoven, Rombouts, Yarkoni, et al., 2015; Richey et al., 2014) and tend to catastrophize social losses (Clark & Wells, 1995; Hofmann, 2007). Our results demonstrated quantitatively that when the perceived social loss overpowers social gain, it can directly lead to avoidant behaviors. Similar to self-efficacy, these results demonstrate a dissociation between beliefs about the social context and the tendency of avoidance, indicating that individuals with strong social loss aversion and/or low social gain sensitivity may require interventions that specifically target this factor in their treatment planning.

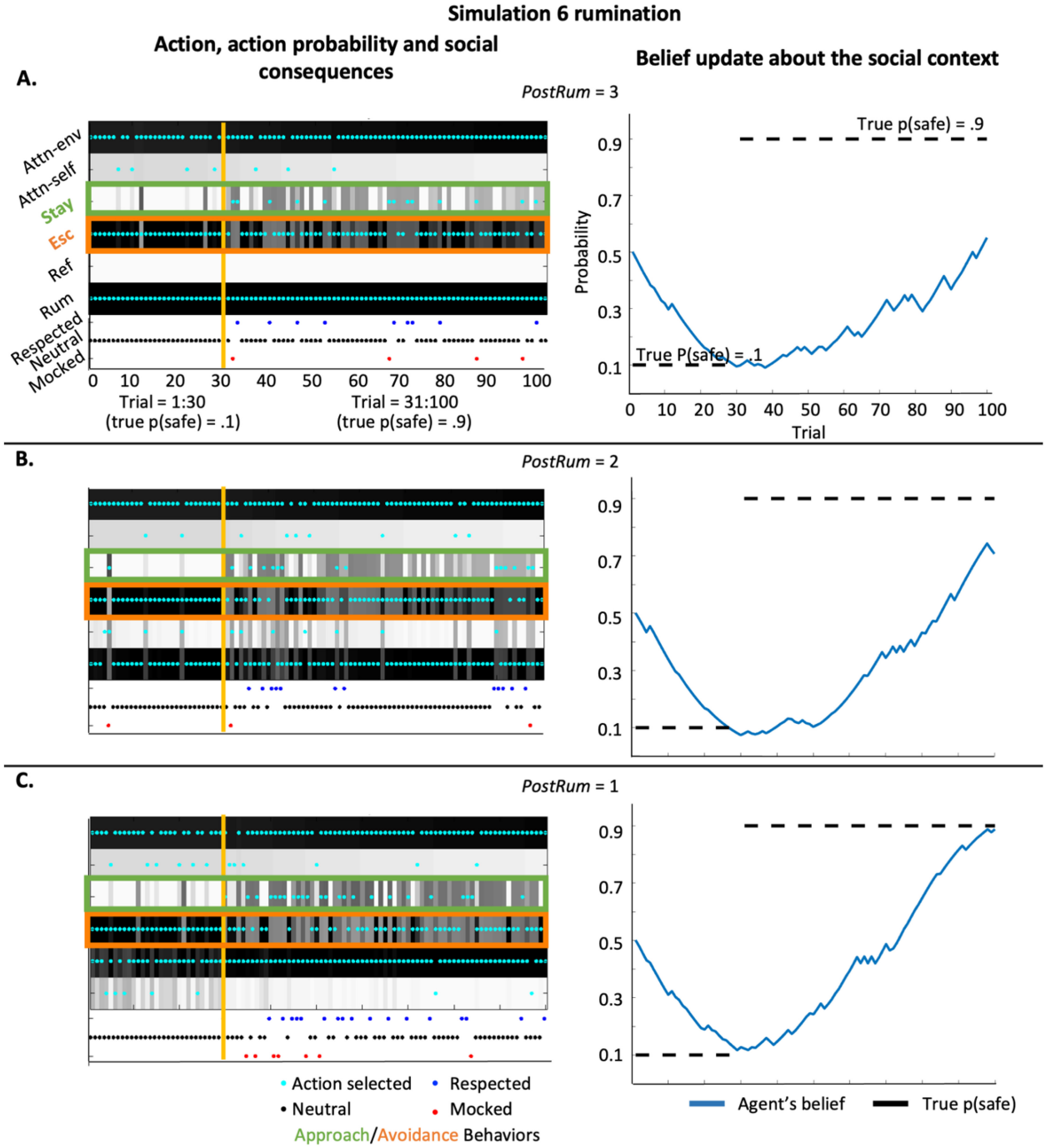
A preexisting biased belief about the social context being threatening showed different impacts compared to the above two factors. Our simulation findings indicate that an initial negative bias alone can be corrected over time, albeit at a slower pace compared to individuals with an optimistic bias (Simulation 2). Importantly, a preexisting negative bias about the social context alone does not appear to significantly induce safety behaviors. This means that at a given moment in life, individuals with negative biases about social contexts are more susceptible to developing negative social appraisals and require more positive experiences to rectify their beliefs (Voncken et al., 2010), but this factor itself is not sufficient to bring about safety behaviors.

A cognitive factor that impaired both social belief update and action choice is self-focused attention, which is commonly observed in individuals with SAD and subclinical symptoms (Piccirillo et al., 2016; Ranta et al., 2014; Thomas et al., 2012). While a moderate level of self-focused attention is not necessarily pathological, high levels of self-focused attention can have negative impacts on social performances, particularly for at-risk populations (Leigh et al., 2021; Spurr & Stopa, 2002). The Simulation 5 clearly demonstrated that heightened self-focused attention leads to inaccurate social beliefs, as individuals rely more on their internal bodily sensations (interoceptive) rather than external cues (exteroceptive) to assess social contexts. In our simulation, when engaged in self-focused attention, the agent completely ignored exteroceptive information. Although we showcased an extreme example to amplify the effect, this outcome did not contradict with clinical presentations of SAD (Norton & Abbott, 2016; Spurr & Stopa, 2002; Woody et al., 1997). Importantly, individuals with self-focused attention alone failed to select the optimal actions across both hostile and supportive environments. The profound impact of attention allocation in our simulation highlights the importance of assessing and treating this cognitive style in clinical practice and may help explain why psychotherapy with cognitive components tend to yield better treatment outcomes than exposure therapy alone (Clark et al., 2003).

Finally, in Simulation 6, rumination did not directly cause SAD symptoms, but exacerbated negative beliefs after encountering undesired consequences. This is consistent with the proposed function of rumination as a critical maintenance factor intensifying anxious feelings and negative self-appraisals (Abbott & Rapee, 2004)

Figure 7

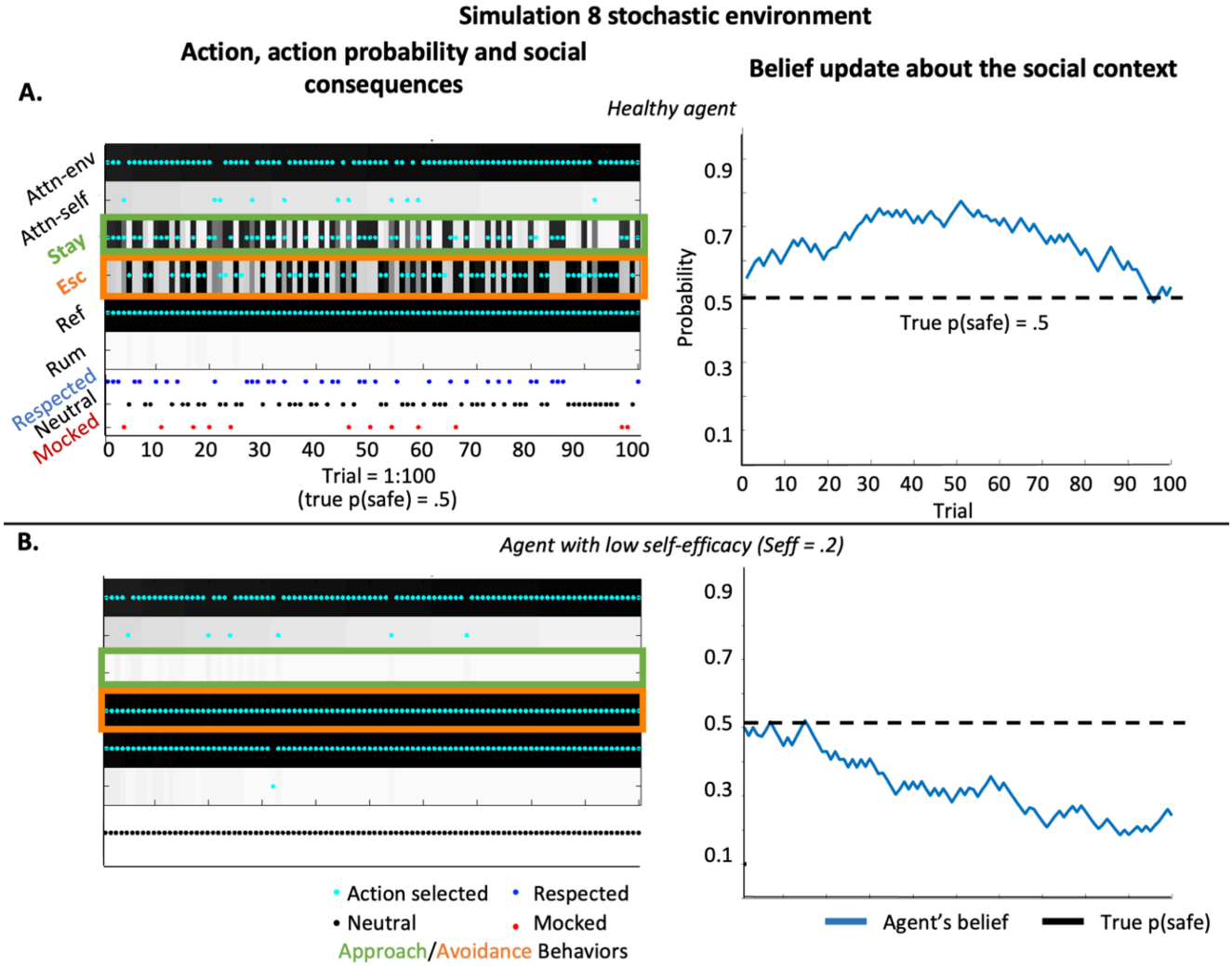
Action Selection, Action Probability, Perceived Social Consequences, and Belief Update for Simulation 6



Note. Results for agents who differed in rumination tendency. (A) PostRum = 3; (B) PostRum = 2; (C) PostRum = 1. Left panels: Action, action probability, and perceived social consequences for agents with different tendency for rumination. As in Figure 2, the yellow (light gray) line indicates the changing point from the hostile (the probability of the context being safe = .1) to the supportive (the probability of the context being safe = .9) social environment. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4, and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

Figure 8

Action Selection, Action Probability, Perceived Social Consequences, and Belief Update for Simulation 8



Note. Results for (A) a healthy agent, and (B) an agent with low self-efficacy in a stochastic social environment. Left panels: Action, action probability, and perceived social consequences for the simulated agents. Different from the previous seven simulations, the probability of the social context being safe was set as .5 throughout the 100 trials. For all panels, darker shades represent higher probability values. Cyan (light gray) dots on the hidden state probability mappings mark the actual states for each trial. Blue (dark gray), black, and red (light gray) dots on the bottom three rows indicating the perceived social consequences indicate socially successful, neutral, and unsuccessful outcomes observed by the agent, respectively. The key behavioral markers, namely the agent's approach/avoidance decisions, were highlighted using green (light gray) and orange (dark gray) colors, respectively. Belief update of the social context for agents with self-efficacy at .2, .4, and .6 is shown on the right panels. Blue (gray) curves depict how the agent's beliefs about the latent social context evolved throughout 100 trials. Black lines mark the true probability of the social context being safe. Attn-env = attending to the environment; Attn-self = attending to self; Esc = escape; Ref = realistic reflecting; Rum = rumination. See the online article for the color version of this figure.

and worsening other maladaptive cognitive processes (Chen et al., 2013).

Simulation 7 showed the clear negative trajectory for people with all of the aforementioned vulnerabilities. Overall, these findings highlight distinctive impacts of these factors on social belief update and safety behavior choice, demonstrating the potential individual differences in symptom development trajectory. Additionally, we included two simulations (Simulation 8 and Supplementary Simulation 2) where we manipulated the true social environment setting to further demonstrate the applicability of the present model across different social environments. Using quantification of hidden computational mechanisms can facilitate the design of personalized treatments, thereby increasing treatment efficacy (Fernandes et al., 2017). These implications are in line with precision medicine and client-centered evidence-based practice (Christon et al., 2015). Moreover, quantitative formalization of such a holistic disorder-specific psychopathology model, especially operationalizing various cognitive factors in context, significantly expanded the insights provided by existing research using active inference to account for isolated factors and simple cognitive–behavioral interactions (Smith et al., 2021).

Active Inference-Based CBT Model Enhances Conceptual Clarity and Embodies Clinical Theories of SAD as Falsifiable Hypotheses

The present active inferenced-based CBT model allows for the formal implementation of various cognitive and behavioral factors. A general benefit of translating psychological constructs from the conceptual model to specific parameters in the quantitative model is that by doing so we circumvented the limitations and variability of verbal expressions (Eaton et al., 2023), enhancing conceptual clarity. While remaining faithful to the Hofmann (2007) model, we made notable revisions and refinements when constructing our computational model, aiming to reduce ambiguity and increase clarity in preexisting theories. For example, we unified "negative self-perception," "low perceived emotional control," and "perceived poor social skills" in the conceptual model with a one single factor of self-efficacy, which influenced the agent's predictions regarding social consequences. Doing so reduced the overlapping between factors and increased the parsimony. However, it is important to note that for future research aiming to differentiate between these perceptual outcomes, one can modify the quantitative model to incorporate these nuances.

Active inference modeling offers a high level of flexibility in examining targeted parameters. In our quantitative CBT model of SAD, various components serve as free parameters to reveal individual differences. Because a large number of free parameters posit potential challenges such as model complexity, overfitting, and difficulties in parameter recovery (Gelman, 2006), only a subset of cognitive–behavioral factors was allowed to vary in our active inference model based on the Hofmann (2007) model, and we refrain from claiming this as the definitive quantitative model of SAD. Rather, we present a demonstrative example that actively integrates the CBT model for psychopathology with the active inference modeling approach. This work showcases the potential for synergizing the practical benefits of clinical and computational psychology. Future research fitting our active inference model of SAD to human data shall consider the parameters most relevant to the research question. Other factors, such as the type and probability of social gains, may regulate pathways to SAD differently (Matyjek et al., 2020), and can be incorporated into the computational model when necessary. Additionally, while we adopted a fixed learning rate in our simulations, deficits in dynamic adaptation could contribute to the maladaptive behaviors (Piray et al., 2019). We did not focus on learning rate, as it was not a major factor in the conceptual model (Heimberg et al., 2010; Hofmann, 2007). However, given the wide influence of learning rate (Caletti et al., 2022; Gillan et al., 2016; Hayes et al., 2006; Lissek et al., 2005), we have included an additional simulation, demonstrating the effects of altered learning rate in the online supplemental materials.

In addition to including different clinical constructs, future studies may manipulate cognitive and behavioral factors in various building blocks in the active inference framework (e.g., allowing the transition matrix ($\boldsymbol{B}$) to vary) or construct different model specifications. In this regard, it is crucial to ensure a balance between parsimony and theoretical coherency, as well as close alignments between computational methodology and research questions (Wilson & Collins, 2019). Building on the current quantitative account of a CBT model for psychopathology, future research could design novel experimental tasks and reasonably combine them with questionnaires and interviews to provide valid and reliable measures for each relevant cognitive–behavioral factor in a holistic manner (Zorowitz & Niv, 2023). It is important to clarify that we do not claim quantitative modeling coupled with simulation can replace the empirical studies using real subjects in model testing. Instead, we stress that (active inference-based) quantitative modeling of complex conceptual models can facilitate theory testing by generating specific predictions, for example, based on thought experiments and simulations, which can be tested against real human and data. Also, such an approach can provide insights which have the potential to advance clinical theories.

In sum, such active inference-based quantitative modeling requires the researcher to build mechanistic models by converting clinical theory descriptions into hypotheses-driven quantitative relationships of clearly defined variables, thus enabling specific predictions. These predictions can be tested by both simulation and human data, facilitating hypothesis falsification.

Synergy Between Active Inference and CBT as a New Approach to Decipher Psychopathology

On a broader scope, the integration of behavior and cognition within the closed loop of active inference provides a natural framework for quantifying CBT conceptualizations of psychopathology. In this symbiotic relationship between CBT and active inference, we emphasize the equal importance of methodological advancement and clinical theories for understanding and effectively addressing mental health problems (Wright & Kaurin, 2020). A multimethod integrative inquiry can help uncover crucial effects that may be overlooked in individual fields (Corlett et al., 2020; Wang & Krystal, 2014). Instead of examining isolated factors as typically done in previous computational modeling works, constructing a holistic CBT model enabled us to understand the interplay between different processes, which could carry significant meanings (Schoen et al., 2011). Through the SAD example, we demonstrate that computational models, informed by well-established clinical theories, can help reveal fine-grained pathways and subtypes of SAD that correspond to the individual differences widely observed in real life.

An important feature of our modeling approach is that the generative model allows us to provide a mechanistic narrative of the agent's psychopathology profile based on the CBT model. As demonstrated in our simulations, the active inference model accommodated hypothetical predisposing factors of an agent in accounting for their behaviors under uncertainty. In other words, having a generative model that can yield observable behaviors means that mechanistic explanations exist for the belief updating process that is, by construction, internally consistent. This approach has the potential to advance our understanding of the specific mechanisms underscoring the pathological decision-making process. The present work serves as a proof-of-concept, translating the key factors of the conceptual CBT model for SAD into a comprehensive quantitative framework. This approach lays the foundation for future investigations of both theoretical conceptualizations and computational interests.

Another key benefit of such generative models is that they can be used to explain empirical data and test hypotheses regarding predispositions. That is, the active inference model proposed in the current article can be applied to explain observed responses in an individual human subject or a cohort of subjects (e.g., Smith et al., 2021).

This corresponds to finding the model parameter values that render the choice behavior most likely (Schwartenbeck & Friston, 2016). Based on the complete class theorem (Brown, 1981; Wald, 1947), for any pair of behaviors and preferences, some prior beliefs exist that render the behavior Bayes optimal in the sense of active inference. In practice, by fitting empirical data and conducting parameter estimation, one can find parameter values that are most likely to produce the observed responses. Moreover, using Bayesian model comparison, both the structure of the model and the parameter values, reflecting hypotheses, can be compared and estimated. Hence, the current approach can be used to achieve computational phenotyping.[5] While previous computational psychiatry research has pursued similar objectives, these efforts have often been less aligned with comprehensive clinical models (Koban et al., 2017, 2023; Smith et al., 2021). Our work bridges the gap between clinical and computational research by offering a platform for future studies to integrate the rigor of computational methods with well-established, clinically valuable conceptual models. This integration, coupled with appropriate experimental paradigms and human data, paves the way for more nuanced and clinically relevant computational approaches in psychiatry.

To further illustrate the practical application of our model, we outline a sample behavioral paradigm, modified based on the social stress induction task (A. P. Allen et al., 2016) (see a demonstration in Figure S4 in the online supplemental materials). In this paradigm, participants are asked in each trial to give a short presentation in front of an audience (confederates). The experiment begins with participants reporting their prior expectations about the upcoming presentation. During the presentation, participants choose between attending to internal bodily sensations or external cues from the audience to inform their decision to continue or discontinue the presentation. After the presentation, participants observe audience responses and reflect on their performance. The scenario is then repeated with different audience groups, allowing for assessment of learning and adaptation. The confederates will provide different expressions and/or comments according to the predetermined social environment. This paradigm enables a systematic assessment on the cognitive–behavioral processes related to SAD as examined in our simulations. Researchers can fit participants' self-reported beliefs and choice data generated by such a behavioral paradigm into our computational model to test hypotheses about the key processes underlying different phenotypes of SAD. As aforementioned, one can test different generative models representing different hypotheses. These models may include different number and combinations of free parameters and/or different prior values of a parameter. By using Bayesian model comparison (functions available through SPM [Penny et al., 2006]), the structure and parameter values of the best-fitting model can then be used to explain the psychopathological process in human subjects. This provides a quantitative method for both personalized modeling of individual differences in psychopathology and testing competing hypotheses of psychopathology. In future work, we plan to illustrate this form of hypothesis testing by proposing a variety of generative models that accommodate competing theories for SAD pathogenesis and compare these models' performance in accounting for empirical data.

The model construction process and mechanistic insights gained from the systematic simulation can be easily generalized to examine other forms of psychopathology where CBT has proven useful, such as depression, generalized anxiety, and obsessive–compulsive disorders. These disorders are all characterized by cognitive biases, misinterpretations, and maladaptive behaviors (Everaert et al., 2018; Kozak & Foa, 1994). Building CBT-based quantitative models holds the potential to better elucidate the computational processes behind the cognition-action dynamics in these psychiatric conditions.

Limitations

Each CBT model can often be applied to a variety of settings, cognitions, and actions. This all-encompassing feature of CBT makes it a flexible and powerful tool in explaining maladaptive behaviors. However, it also posts challenges when constructing concrete quantitative models. In our SAD demonstration, the concreteness is partially achieved by using exemplary descriptors (e.g., "threatening" vs. "safe" social setting). Employing representative descriptors serve the purpose of providing the first example of using active inference to quantify a holistic CBT model. Future research, building on our humble example, can expand on the complexity and nuances, for example, specifying the social setting and action options with higher granularity. As aforementioned, the specific model construction should consider the purpose and experimental design of the study the model is concerned with. In addition, for the purpose of the present article, we used the Hofmann (2007) as the target conceptual model, given its comprehensiveness and relative recency. The same quantitative formalization process using active inference can be applied to other CBT models of SAD (e.g., the Clark and Wells model [Clark & Wells, 1995]), and the interpretations may vary depending on the targeted conceptual model. Another direction for future research is to investigate the neuropsychological processes linked to SAD by incorporating relevant biological factors into the model. Some of these biological factors have been included in the active inference framework (e.g., the role of dopamine in minimizing free energy [FitzGerald et al., 2015]).

Concluding Remarks

To conclude, we presented a holistic quantitative model of SAD by integrating CBT and active inference modeling. We showed the mechanistic pathways by linking a range of cognitive and behavioral factors to the emergence and maintenance of SAD symptoms and quantified their relative contribution. Our approach not only motivates reformulation and advancement of psychopathology conceptualizations from a theoretic perspective, but also highlights potential therapeutic targets for more personalized, effective intervention in clinical practice.

[5] We thank an anonymous reviewer for suggesting the content on computational phenotyping. The reviewer also provided the relevant background knowledge.

References

Abbott, M. J., & Rapee, R. M. (2004). Post-event rumination and negative self-appraisal in social phobia before and after treatment. *Journal of Abnormal Psychology*, *113*(1), 136–144. https://doi.org/10.1037/0021-843X.113.1.136

Allen, A. P., Kennedy, P. J., Dockray, S., Cryan, J. F., Dinan, T. G., & Clarke, G. (2016). The trier social stress test: Principles and practice. *Neurobiology of Stress*, *6*, 113–126. https://doi.org/10.1016/j.ynstr.2016.11.001

Allen, M., Levy, A., Parr, T., & Friston, K. J. (2022). In the body's eye: The computational anatomy of interoceptive inference. *PLOS Computational Biology*, *18*(9), Article e1010490. https://doi.org/10.1371/journal.pcbi.1010490

Alonso, J., Angermeyer, M. C., Bernert, S., Bruffaerts, R., Brugha, T. S., Bryson, H., de Girolamo, G., Graaf, R., Demyttenaere, K., Gasquet, I., Haro, J. M., Katz, S. J., Kessler, R. C., Kovess, V., Lépine, J. P., Ormel, J., Polidori, G., Russo, L. J., Vilagut, G., … ESEMeD/MHEDEA 2000 Investigators, European Study of the Epidemiology of Mental Disorders (ESEMeD) Project. (2004). Disability and quality of life impact of mental disorders in Europe: Results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatrica Scandinavica. Supplementum*, *109*(s420), 38–46. https://doi.org/10.1111/j.1600-0047.2004.00329.x

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Andino, M. V., Km, G., & Ja, R. (2024). Can dialectical behavior therapy skills group treat social anxiety disorder? A brief integrative review. *Frontiers in Psychology*, *14*, Article 1331200. https://doi.org/10.3389/fpsyg.2023.1331200

Bandura, A., Caprara, G. V., Barbaranelli, C., Gerbino, M., & Pastorelli, C. (2003). Role of affective self-regulatory efficacy in diverse spheres of psychosocial functioning. *Child Development*, *74*(3), 769–782. https://doi.org/10.1111/1467-8624.00567

Beck, J. S. (2021). *Cognitive behavior therapy: Basics and beyond* (3rd ed., pp. xvi, 414). The Guilford Press.

Biehl, M., Guckelsberger, C., Salge, C., Smith, S. C., & Polani, D. (2018). Expanding the active inference landscape: More intrinsic motivations in the perception-action loop. *Frontiers in Neurorobotics*, *12*, Article 45. https://doi.org/10.3389/fnbot.2018.00045

Binelli, C., Muñiz, A., Sanches, S., Ortiz, A., Navines, R., Egmond, E., Udina, M., Batalla, A., López-Sola, C., Crippa, J. A., Subirà, S., & Martín-Santos, R. (2015). New evidence of heterogeneity in social anxiety disorder: Defining two qualitatively different personality profiles taking into account clinical, environmental and genetic factors. *European Psychiatry*, *30*(1), 160–165. https://doi.org/10.1016/j.eurpsy.2014.09.418

Bögels, S. M., & Mansell, W. (2004). Attention processes in the maintenance and treatment of social phobia: Hypervigilance, avoidance and self-focused attention. *Clinical Psychology Review*, *24*(7), 827–856. https://doi.org/10.1016/j.cpr.2004.06.005

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488. https://doi.org/10.1016/j.tics.2012.08.006

Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics*, *9*(6), 1289–1300. https://doi.org/10.1214/aos/1176345645

Brozovich, F., & Heimberg, R. G. (2008). An analysis of post-event processing in social anxiety disorder. *Clinical Psychology Review*, *28*(6), 891–903. https://doi.org/10.1016/j.cpr.2008.01.002

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review*, *26*(1), 17–31. https://doi.org/10.1016/j.cpr.2005.07.003

Caletti, E., Massimo, C., Magliocca, S., Moltrasio, C., Brambilla, P., & Delvecchio, G. (2022). The role of the acceptance and commitment therapy in the treatment of social anxiety: An updated scoping review. *Journal of Affective Disorders*, *310*, 174–182. https://doi.org/10.1016/j.jad.2022.05.008

Caouette, J. D., & Guyer, A. E. (2014). Gaining insight into adolescent vulnerability for social anxiety from developmental cognitive neuroscience. *Developmental Cognitive Neuroscience*, *8*, 65–76. https://doi.org/10.1016/j.dcn.2013.10.003

Carlbring, P., Apelstrand, M., Sehlin, H., Amir, N., Rousseau, A., Hofmann, S. G., & Andersson, G. (2012). Internet-delivered attention bias modification training in individuals with social anxiety disorder—A double blind randomized controlled trial. *BMC Psychiatry*, *12*(1), Article 66. https://doi.org/10.1186/1471-244X-12-66

Chen, J., Wang, Z., Wu, Y., Cai, Y., Shen, Y., Wang, L., & Shi, S. (2013). Differential attentional bias in generalized anxiety disorder and panic disorder. *Neuropsychiatric Disease and Treatment*, *9*, 73–80. https://doi.org/10.2147/NDT.S36822

Christon, L. M., McLeod, B. D., & Jensen-Doss, A. (2015). Evidence-based assessment meets evidence-based treatment: An approach to science-informed case conceptualization. *Cognitive and Behavioral Practice*, *22*(1), 36–48. https://doi.org/10.1016/j.cbpra.2013.12.004

Clark, D. M., Ehlers, A., McManus, F., Hackmann, A., Fennell, M., Campbell, H., Flower, T., Davenport, C., & Louis, B. (2003). Cognitive therapy versus fluoxetine in generalized social phobia: A randomized placebo-controlled trial. *Journal of Consulting and Clinical Psychology*, *71*(6), 1058–1067. https://doi.org/10.1037/0022-006X.71.6.1058

Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. In R. G. Heimberg, M. R. Liebowitz, D. A. Hope, & F. R. Schneier (Eds.), *Social phobia: Diagnosis, assessment, and treatment* (pp. 69–93). The Guilford Press.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., & Friston, K. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, *10*, Article 679. https://doi.org/10.3389/fpsyg.2019.00679

Corlett, P. R., Mohanty, A., & MacDonald, A. W. (2020). What we think about when we think about predictive processing. *Journal of Abnormal Psychology*, *129*(6), 529–533. https://doi.org/10.1037/abn0000632

Craske, M. G. (2017). *Cognitive-behavioral therapy* (2nd ed.). American Psychological Association. https://doi.org/10.1037/0000027-000

Cremers, H. R., Veer, I. M., Spinhoven, P., Rombouts, S. A. R. B., & Roelofs, K. (2015). Neural sensitivity to social reward and punishment anticipation in social anxiety disorder. *Frontiers in Behavioral Neuroscience*, *8*, Article 439. https://doi.org/10.3389/fnbeh.2014.00439

Cremers, H. R., Veer, I. M., Spinhoven, P., Rombouts, S. A. R. B., Yarkoni, T., Wager, T. D., & Roelofs, K. (2015). Altered cortical-amygdala coupling in social anxiety disorder during the anticipation of giving a public speech. *Psychological Medicine*, *45*(7), 1521–1529. https://doi.org/10.1017/S0033291714002657

Cuijpers, P. (2019). Targets and outcomes of psychotherapies for mental disorders: An overview. *World Psychiatry*, *18*(3), 276–285. https://doi.org/10.1002/wps.20661

Eaton, N. R., Bringmann, L. F., Elmer, T., Fried, E. I., Forbes, M. K., Greene, A. L., Krueger, R. F., Kotov, R., McGorry, P. D., Mei, C., & Waszczuk, M. A. (2023). A review of approaches and models in psychopathology conceptualization research. *Nature Reviews Psychology*, *2*(10), 622–636. https://doi.org/10.1038/s44159-023-00218-4

Enock, P. M., Hofmann, S. G., & McNally, R. J. (2014). Attention bias modification training via smartphone to reduce social anxiety: A randomized, controlled multi-session experiment. *Cognitive Therapy and Research*, *38*(2), 200–216. https://doi.org/10.1007/s10608-014-9606-z

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking through tinted glasses: Depression and social anxiety are related to both interpretation biases and inflexible negative interpretations. *Clinical Psychological Science*, *6*(4), 517–528. https://doi.org/10.1177/2167702617747968

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of "precision psychiatry." *BMC Medicine*, *15*(1), Article 80. https://doi.org/10.1186/s12916-017-0849-x

Fink, M., Akimova, E., Spindelegger, C., Hahn, A., Lanzenberger, R., & Kasper, S. (2009). Social anxiety disorder: Epidemiology, biology and treatment. *Psychiatria Danubina*, *21*(4), 533–542.

FitzGerald, T. H. B., Dolan, R. J., & Friston, K. (2015). Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience*, *9*, Article 136. https://doi.org/10.3389/fncom.2015.00136

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. (2023). Computational psychiatry: From synapses to sentience. *Molecular Psychiatry*, *28*(1), 256–268. https://doi.org/10.1038/s41380-022-01743-z

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053

Gaudiano, B. A., & Herbert, J. D. (2003). Preliminary psychometric evaluation of a new self-efficacy scale and its relationship to treatment outcome in social anxiety disorder. *Cognitive Therapy and Research*, *27*(5), 537–555. https://doi.org/10.1023/A:1026355004548

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gerrans, P., & Murray, R. J. (2020). Interoceptive active inference and self-representation in social anxiety disorder (SAD): Exploring the neurocognitive traits of the SAD self. *Neuroscience of Consciousness*, *2020*(1), Article niaa026. https://doi.org/10.1093/nc/niaa026

Gillan, C. M., Robbins, T. W., Sahakian, B. J., van den Heuvel, O. A., & van Wingen, G. (2016). The role of habit in compulsivity. *European Neuropsychopharmacology*, *26*(5), 828–840. https://doi.org/10.1016/j.euroneuro.2015.12.033

Grant, B. F., Hasin, D. S., Blanco, C., Stinson, F. S., Chou, S. P., Goldstein, R. B., Dawson, D. A., Smith, S., Saha, T. D., & Huang, B. (2005). The epidemiology of social anxiety disorder in the United States: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *The Journal of Clinical Psychiatry*, *66*(11), 1351–1361. https://doi.org/10.4088/jcp.v66n1102

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, *5*(4), 269–306. https://doi.org/10.1016/j.cogsys.2004.02.002

Gregory, B., Peters, L., & Rapee, R. M. (2016). The self in social anxiety. In G. Doron, M. Nedeljkovic, M. Mikulincer, M. Kyrios, R. Moulding, & S. S. Bhar (Eds.), *The self in understanding and treating psychological disorders* (pp. 91–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139941297.011

Harrison, A. W., Rainer, R. K., Hochwarter, W. A., & Thompson, K. R. (1997). Testing the self-efficacy—Performance linkage of social—Cognitive theory. *The Journal of Social Psychology*, *137*(1), 79–87. https://doi.org/10.1080/00224549709595415

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*(6), 930–957. https://doi.org/10.1037/met0000303

Hauke, G. (2016). *European Psychotherapy 2016/2017: Embodiment in psychotherapy*. BoD—Books on Demand.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour Research and Therapy*, *44*(1), 1–25. https://doi.org/10.1016/j.brat.2005.06.006

Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2010). Chapter 15—A cognitive behavioral model of social anxiety disorder: Update and extension. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety* (2nd ed., pp. 395–422). Academic Press. https://doi.org/10.1016/B978-0-12-375096-9.00015-8

Hofmann, S. G. (2007). Cognitive factors that maintain social anxiety disorder: A comprehensive model and its treatment implications. *Cognitive Behaviour Therapy*, *36*(4), 193–209. https://doi.org/10.1080/16506070701421313

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1

Hofmann, S. G., & Otto, M. W. (2008). *Cognitive behavioral therapy for social anxiety disorder: Evidence-based and disorder-specific treatment techniques*. Routledge. https://doi.org/10.4324/9780203927526

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Holmes, E., Parr, T., Griffiths, T. D., & Friston, K. J. (2021). Active inference, selective attention, and the cocktail party problem. *Neuroscience & Biobehavioral Reviews*, *131*, 1288–1304. https://doi.org/10.1016/j.neubiorev.2021.09.038

Hudd, T., & Moscovitch, D. A. (2022). Reconnecting in the face of exclusion: Individuals with high social anxiety may feel the push of social pain, but not the pull of social rewards. *Cognitive Therapy and Research*, *46*(2), 420–435. https://doi.org/10.1007/s10608-021-10263-z

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(1), 3–19. https://doi.org/10.1038/s41386-020-0746-4

Iancu, I., Bodner, E., & Ben-Zion, I. Z. (2015). Self esteem, dependency, self-efficacy and self-criticism in social anxiety disorder. *Comprehensive Psychiatry*, *58*, 165–171. https://doi.org/10.1016/j.comppsych.2014.11.018

Javaid, S. F., Hashim, I. J., Hashim, M. J., Stip, E., Samad, M. A., & Ahbabi, A. A. (2023). Epidemiology of anxiety disorders: Global burden and sociodemographic associations. *Middle East Current Psychiatry*, *30*(1), Article 44. https://doi.org/10.1186/s43045-023-00315-3

Jin, J., Jonas, K., & Mohanty, A. (2023). Linking the past to the future by predictive processing: Implications for psychopathology. *Journal of Psychopathology and Clinical Science*, *132*(3), 249–262. https://doi.org/10.1037/abn0000730

Joormann, J., Dkane, M., & Gotlib, I. H. (2006). Adaptive and maladaptive components of rumination? Diagnostic specificity and relation to depressive biases. *Behavior Therapy*, *37*(3), 269–280. https://doi.org/10.1016/j.beth.2006.01.002

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*, *112*(4), 323–343. https://doi.org/10.1007/s00422-018-0753-2

Kent, J. S., Markon, K., & MacDonald, A. W. (2023). Theories of psychopathology: Introduction to a special section. *Journal of Psychopathology and Clinical Science*, *132*(3), 223–227. https://doi.org/10.1037/abn0000824

Kiverstein, J., Rietveld, E., Slagter, H. A., & Denys, D. (2019). Obsessive compulsive disorder: A pathology of self-confidence? *Trends in Cognitive Sciences*, *23*(5), 369–372. https://doi.org/10.1016/j.tics.2019.02.005

Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D., & Arch, J. J. (2023). Brain mediators of biased social learning of self-perception in social anxiety disorder. *Translational Psychiatry*, *13*(1), Article 292. https://doi.org/10.1038/s41398-023-02587-z

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., & Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion (Washington, D.C.)*, *17*(8), 1144–1155. https://doi.org/10.1037/emo0000296

Kocovski, N. L., & Rector, N. A. (2007). Predictors of post-event rumination related to social anxiety. *Cognitive Behaviour Therapy*, *36*(2), 112–122. https://doi.org/10.1080/16506070701232090

Kopala-Sibley, D. C., Zuroff, D. C., Russell, J. J., & Moskowitz, D. S. (2014). Understanding heterogeneity in social anxiety disorder: Dependency and self-criticism moderate fear responses to interpersonal cues. *British Journal of Clinical Psychology*, *53*(2), 141–156. https://doi.org/10.1111/bjc.12032

Kozak, M. J., & Foa, E. B. (1994). Obsessions, overvalued ideas, and delusions in obsessive-compulsive disorder. *Behaviour Research and Therapy*, *32*(3), 343–353. https://doi.org/10.1016/0005-7967(94)90132-5

Leary, M. R., & Atherton, S. C. (1986). Self-efficacy, social anxiety, and inhibition in interpersonal encounters. *Journal of Social and Clinical Psychology*, *4*(3), 256–267. https://doi.org/10.1521/jscp.1986.4.3.256

Leary, M. R., & Kowalski, R. M. (1997). *Social anxiety*. Guilford Press.

Leigh, E., Chiu, K., & Clark, D. M. (2021). Self-focused attention and safety behaviours maintain social anxiety in adolescents: An experimental study. *PLOS ONE*, *16*(2), Article e0247703. https://doi.org/10.1371/journal.pone.0247703

Leon, A. C., Portera, L., & Weissman, M. M. (1995). The social costs of anxiety disorders. *The British Journal of Psychiatry*, *166*(Suppl. 27), 19–22. https://doi.org/10.1192/S0007125000293355

Limongi, R., Bohaterewicz, B., Nowicka, M., Plewka, A., & Friston, K. J. (2018). Knowing when to stop: Aberrant precision and evidence accumulation in schizophrenia. *Schizophrenia Research*, *197*, 386–391. https://doi.org/10.1016/j.schres.2017.12.018

Linson, A., & Friston, K. (2019). Reframing PTSD for computational psychiatry with the active inference framework. *Cognitive Neuropsychiatry*, *24*(5), 347–368. https://doi.org/10.1080/13546805.2019.1665994

Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behaviour Research and Therapy*, *43*(11), 1391–1424. https://doi.org/10.1016/j.brat.2004.10.007

Mace, F. C., Lalli, J. S., & Lalli, E. P. (1991). Functional analysis and treatment of aberrant behavior. *Research in Developmental Disabilities*, *12*(2), 155–180. https://doi.org/10.1016/0891-4222(91)90004-C

Mansell, W., Clark, D. M., & Ehlers, A. (2003). Internal versus external attention in social anxiety: An investigation using a novel paradigm. *Behaviour Research and Therapy*, *41*(5), 555–572. https://doi.org/10.1016/S0005-7967(02)00029-3

Matyjek, M., Meliss, S., Dziobek, I., & Murayama, K. (2020). A multidimensional view on social and non-social rewards. *Frontiers in Psychiatry*, *11*, Article 818. https://doi.org/10.3389/fpsyt.2020.00818

Meehl, P. E. (1977). Specific etiology and other forms of strong influence: Some quantitative meanings. *The Journal of Medicine and Philosophy*, *2*(1), 33–53. https://doi.org/10.1093/jmp/2.1.33

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mellings, T. M. B., & Alden, L. E. (2000). Cognitive processes in social anxiety: The effects of self-focus, rumination and anticipatory processing. *Behaviour Research and Therapy*, *38*(3), 243–257. https://doi.org/10.1016/S0005-7967(99)00040-6

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus, negative life events, and negative affect. *Behaviour Research and Therapy*, *46*(9), 1034–1039. https://doi.org/10.1016/j.brat.2008.06.004

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin*, *128*(4), 638–662. https://doi.org/10.1037/0033-2909.128.4.638

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. https://doi.org/10.1037//0022-3514.61.1.115

Norton, A. R., & Abbott, M. J. (2016). Self-focused cognition in social anxiety: A review of the theoretical and empirical literature. *Behaviour Change*, *33*(1), 44–64. https://doi.org/10.1017/bec.2016.2

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2006). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier Science & Technology. https://ebookcentral.proquest.com/lib/hkuhk/detail.action?docID=282095

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Piccirillo, M. L., Taylor Dryman, M., & Heimberg, R. G. (2016). Safety behaviors in adults with social anxiety: Review and future directions. *Behavior Therapy*, *47*(5), 675–687. https://doi.org/10.1016/j.beth.2015.11.005

Pineles, S., & Mineka, S. (2005). Attentional biases to internal and external sources of potential threat in social anxiety. *Journal of Abnormal Psychology*, *114*(2), 314–318. https://doi.org/10.1037/0021-843X.114.2.314

Piray, P., Ly, V., Roelofs, K., Cools, R., & Toni, I. (2019). Emotionally aversive cues suppress neural systems underlying optimal learning in socially anxious individuals. *Journal of Neuroscience*, *39*(8), 1445–1456. https://doi.org/10.1523/JNEUROSCI.1394-18.2018

Powers, M. B., de Kleine, R. A., & Smits, J. A. J. (2017). Core mechanisms of cognitive behavioral therapy for anxiety and depression: A review. *Psychiatric Clinics of North America*, *40*(4), 611–623. https://doi.org/10.1016/j.psc.2017.08.010

Rachman, S., Grüter-Andrew, J., & Shafran, R. (2000). Post-event processing in social anxiety. *Behaviour Research and Therapy*, *38*(6), 611–617. https://doi.org/10.1016/s0005-7967(99)00089-3

Ranta, K., Tuomisto, M. T., Kaltiala-Heino, R., Rantanen, P., & Marttunen, M. (2014). Cognition, imagery and coping among adolescents with social anxiety and phobia: Testing the Clark and Wells model in the population. *Clinical Psychology & Psychotherapy*, *21*(3), 252–263. https://doi.org/10.1002/cpp.1833

Rapee, R. M., & Heimberg, R. G. (1997). A cognitive-behavioral model of anxiety in social phobia. *Behaviour Research and Therapy*, *35*(8), 741–756. https://doi.org/10.1016/s0005-7967(97)00022-3

Reilly, E. E., Whitton, A. E., Pizzagalli, D. A., Rutherford, A. V., Stein, M. B., Paulus, M. P., & Taylor, C. T. (2020). Diagnostic and dimensional evaluation of implicit reward learning in social anxiety disorder and major depression. *Depression and Anxiety*, *37*(12), 1221–1230. https://doi.org/10.1002/da.23081

Rheingold, A. A., Herbert, J. D., & Franklin, M. E. (2003). Cognitive bias in adolescents with social anxiety disorder. *Cognitive Therapy and Research*, *27*(6), 639–655. https://doi.org/10.1023/A:1026399627766

Richards, J. M., Patel, N., Daniele-Zegarelli, T., MacPherson, L., Lejuez, C. W., & Ernst, M. (2015). Social anxiety, acute social stress, and reward parameters interact to predict risky decision-making among adolescents. *Journal of Anxiety Disorders*, *29*, 25–34. https://doi.org/10.1016/j.janxdis.2014.10.001

Richey, J. A., Brewer, J. A., Sullivan-Toole, H., Strege, M. V., Kim-Spoon, J., White, S. W., & Ollendick, T. H. (2019). Sensitivity shift theory: A developmental model of positive affect and motivational deficits in social anxiety disorder. *Clinical Psychology Review*, *72*, Article 101756. https://doi.org/10.1016/j.cpr.2019.101756

Richey, J. A., Rittenberg, A., Hughes, L., Damiano, C. R., Sabatino, A., Miller, S., Hanna, E., Bodfish, J. W., & Dichter, G. S. (2014). Common and distinct neural features of social and non-social reward processing in

autism and social anxiety disorder. *Social Cognitive and Affective Neuroscience*, *9*(3), 367–377. https://doi.org/10.1093/scan/nss146

Roth, D. A., & Heimberg, R. G. (2001). Cognitive-behavioral models of social anxiety disorder. *Psychiatric Clinics of North America*, *24*(4), 753–771. https://doi.org/10.1016/S0193-953X(05)70261-6

Rudy, B. M., Davis, T. E., & Matthews, R. A. (2012). The relationship among self-efficacy, negative self-referent cognitions, and social anxiety in children: A multiple mediator model. *Behavior Therapy*, *43*(3), 619–628. https://doi.org/10.1016/j.beth.2011.11.003

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, *33*(3), 674–712. https://doi.org/10.1162/neco_a_01357

Schneider, R. L., Arch, J. J., & Wolitzky-Taylor, K. B. (2015). The state of personalized treatment for anxiety disorders: A systematic review of treatment moderators. *Clinical Psychology Review*, *38*, 39–54. https://doi.org/10.1016/j.cpr.2015.02.004

Schoen, J. L., DeSimone, J. A., & James, L. R. (2011). Exploring joint variance between independent variables and a criterion: Meaning, effect, and size. *Organizational Research Methods*, *14*(4), 674–695. https://doi.org/10.1177/1094428110381787

Schultz, L. T., & Heimberg, R. G. (2008). Attentional focus in social anxiety disorder: Potential for interactive processes. *Clinical Psychology Review*, *28*(7), 1206–1221. https://doi.org/10.1016/j.cpr.2008.04.003

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, *4*, Article 710. https://doi.org/10.3389/fpsyg.2013.00710

Schwartenbeck, P., & Friston, K. (2016). Computational phenotyping in psychiatry: A worked example. *eNeuro*, *3*(4), Article ENEURO.0049-16.2016. https://doi.org/10.1523/ENEURO.0049-16.2016

Sharot, T. (2011). *The optimism bias: A tour of the irrationally positive brain* (pp. xvii, 245). Pantheon/Random House.

Shin, J. Y., Kim, C., & Hwang, H. J. (2022). Prior preference learning from experts: Designing a reward with active inference. *Neurocomputing*, *492*, 508–515. https://doi.org/10.1016/j.neucom.2021.12.042

Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, *107*, Article 102632. https://doi.org/10.1016/j.jmp.2021.102632

Smith, R., Moutoussis, M., & Bilek, E. (2021). Simulating the computational mechanisms of cognitive and behavioral psychotherapeutic interventions: Insights from active inference. *Scientific Reports*, *11*(1), Article 10128. https://doi.org/10.1038/s41598-021-89047-0

Spurr, J. M., & Stopa, L. (2002). Self-focused attention in social phobia and social anxiety. *Clinical Psychology Review*, *22*(7), 947–975. https://doi.org/10.1016/S0272-7358(02)00107-1

Stein, D. J., Lim, C. C. W., Roest, A. M., de Jonge, P., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Benjet, C., Bromet, E. J., Bruffaerts, R., de Girolamo, G., Florescu, S., Gureje, O., Haro, J. M., Harris, M. G., He, Y., Hinkov, H., Horiguchi, I., Hu, C., … WHO World Mental Health Survey Collaborators. (2017). The cross-national epidemiology of social anxiety disorder: Data from the World Mental Health Survey Initiative. *BMC Medicine*, *15*(1), Article 143. https://doi.org/10.1186/s12916-017-0889-2

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92. https://doi.org/10.1016/j.conb.2013.12.007

Tahmassian, K., & Jalali Moghadam, N. (2011). Relationship between self-efficacy and symptoms of anxiety, depression, worry and social avoidance in a normal sample of students. *Iranian Journal of Psychiatry and Behavioral Sciences*, *5*(2), 91–98.

Thomas, S. A., Daruwala, S. E., Goepel, K. A., & De Los Reyes, A. (2012). Using the subtle avoidance frequency examination in adolescent social anxiety assessments. *Child & Youth Care Forum*, *41*(6), 547–559. https://doi.org/10.1007/s10566-012-9181-y

Thomasson, P., & Psouni, E. (2010). Social anxiety and related social impairment are linked to self-efficacy and dysfunctional coping. *Scandinavian Journal of Psychology*, *51*(2), 171–178. https://doi.org/10.1111/j.1467-9450.2009.00731.x

Voncken, M. J., Dijk, C., de Jong, P. J., & Roelofs, J. (2010). Not self-focused attention but negative beliefs affect poor social performance in social anxiety: An investigation of pathways in the social anxiety–social rejection relationship. *Behaviour Research and Therapy*, *48*(10), 984–991. https://doi.org/10.1016/j.brat.2010.06.004

Wald, A. (1947). An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, *18*(4), 549–555. https://doi.org/10.1214/aoms/1177730345

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. https://doi.org/10.1016/j.neuron.2014.10.018

Weinstein, N. D. (1989). Optimistic biases about personal risks. *Science*, *246*(4935), 1232–1233. https://doi.org/10.1126/science.2686031

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

Woods, B. K., Sauer-Zavala, S., Farchione, T. J., & Barlow, D. H. (2020). Isolating the effects of mindfulness training across anxiety disorder diagnoses in the unified protocol. *Behavior Therapy*, *51*(6), 972–983. https://doi.org/10.1016/j.beth.2020.01.001

Woody, S. R., Chambless, D. L., & Glass, C. R. (1997). Self-focused attention in the treatment of social phobia. *Behaviour Research and Therapy*, *35*(2), 117–129. https://doi.org/10.1016/S0005-7967(96)00084-8

Wright, A. G. C., & Kaurin, A. (2020). Integrating structure and function in conceptualizing and assessing pathological traits. *Psychopathology*, *53*(3–4), 189–197. https://doi.org/10.1159/000507590

Wright, A. G. C., Pincus, A. L., & Hopwood, C. J. (2023). Contemporary integrative interpersonal theory: Integrating structure, dynamics, temporal scale, and levels of analysis. *Journal of Psychopathology and Clinical Science*, *132*(3), 263–276. https://doi.org/10.1037/abn0000741

Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 49–74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004

Received June 26, 2024
Revision received October 29, 2024
Accepted November 6, 2024 ■