computational psychiatry

A Computational Model of Hopelessness and Active-Escape Bias in Suicidality

RESEARCH ARTICLE

POVILAS KARVELIS
ANDREEA O. DIACONESCU
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Currently, psychiatric practice lacks reliable predictive tools and a sufficiently detailed mechanistic understanding of suicidal thoughts and behaviors (STB) to provide timely and personalized interventions. Developing computational models of STB that integrate across behavioral, cognitive and neural levels of analysis could help better understand STB vulnerabilities and guide personalized interventions. To that end, we present a computational model based on the active inference framework. With this model, we show that several STB risk markers – hopelessness, Pavlovian bias and active-escape bias – are interrelated via the drive to maximize one's model evidence. We propose four ways in which these effects can arise: (1) increased learning from aversive outcomes, (2) reduced belief decay in response to unexpected outcomes, (3) increased stress sensitivity and (4) reduced sense of stressor controllability. These proposals stem from considering the neurocircuits implicated in STB: how the locus coeruleus – norepinephrine (LC-NE) system together with the amygdala (Amy), the dorsal prefrontal cortex (dPFC) and the anterior cingulate cortex (ACC) mediate learning in response to acute stress and volatility as well as how the dorsal raphe nucleus – serotonin (DRN-5-HT) system together with the ventromedial prefrontal cortex (vmPFC) mediate stress reactivity based on perceived stressor controllability. We validate the model by simulating performance in an Avoid/Escape Go/No-Go task replicating recent behavioral findings. This serves as a proof of concept and provides a computational hypothesis space that can be tested empirically and be used to distinguish planful versus impulsive STB subtypes. We discuss the relevance of the proposed model for treatment response prediction, including pharmacotherapy and psychotherapy, as well as sex differences as it relates to stress reactivity and suicide risk.

CORRESPONDING AUTHOR:
Povilas Karvelis
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada
povilas.karvelis@camh.ca

KEYWORDS:
suicidality; hopelessness; Pavlovian bias; active-escape bias; active inference; computational modelling; simulation study

TO CITE THIS ARTICLE:
Karvelis, P., & Diaconescu, A. O. (2022). A Computational Model of Hopelessness and Active-Escape Bias in Suicidality. *Computational Psychiatry*, 6(1), pp. 34–59. DOI: *https://doi.org/10.5334/cpsy.80*

DOI: 10.5334/cpsy.80

1 INTRODUCTION

Suicide is the second leading cause of death among young adults and among the top ten causes of death across all ages worldwide (Naghavi et al., 2017). Despite decades of research seeking to identify risk factors of suicidal thoughts and behaviors (STB), their predictive ability remains limited (Large et al., 2016; Franklin et al., 2017). Some of the main risk factors include the following: prior psychiatric diagnosis, treatment history, family history of psychopathology, prior self-injurious thoughts and behaviors, substance use and psychosocial stress. However, multivariate suicide risk models based on these factors do not have sufficient sensitivity and specificity in predicting suicide and, even more importantly, lack mechanistic insight to offer clinically useful guidance on selecting optimal individualized interventions (Kessler et al., 2020). As a result, current clinical practice is in need of objective and reliable measures of suicide risk to not have to rely on self-reports, with ~50% of adults not disclosing their suicidal thoughts and remaining invisible for suicide prevention efforts (Mérelle et al., 2018).

In recent years, cognitive theories have proposed several explanations for the progression from emotional distress to suicidal ideation, and to suicide attempts (Van Orden et al., 2010; Klonsky and May, 2015; O'Connor and Kirtley, 2018; Bryan et al., 2020). At the core of these proposals is the recognition that suicide can be viewed as a means to escape mental pain (psychache) (Baumeister, 1990; Verrocchio et al., 2016). While mental pain and hopelessness contribute to suicidal ideation, other factors, collectively termed 'acquired capability for suicide' (e.g., increased physical pain tolerance, access to lethal means), mediate the transition from ideation to suicide attempt (for a review see Klonsky et al. (2018)). While providing useful high-level insights into the different psychological and environmental factors associated with suicidality, the verbal nature of these theories limits their predictive power (Millner et al., 2020; Meehl, 1990). Natural language is inherently vague resulting in intercorrelated constructs on which the theories rest, making it difficult to corroborate or refute them (Millner et al., 2020; Meehl, 1990). This calls for formal theories of suicidality which can be expressed computationally and which can define these constructs operationally (Millner et al., 2020; Dombrovski and Hallquist, 2021). Computational models could allow for a quantification of suicide risk and offer a more mechanistic insight for developing personalized clinical interventions (Nair et al., 2020; Millner et al., 2020). Just as importantly, computational models can help bridge different levels of analysis and establish mechanistic links between behavioral, cognitive, neural and even genetic variables, offering a more integrated understanding of the factors underlying vulnerability to STB (Huys et al., 2021).

One principled way of building such models is to investigate vulnerability to STB through the lens of normative theories of learning and decision making in computational neuroscience (Dombrovski and Hallquist, 2021, 2017). Collectively, STB has been associated with deficits in cognitive control (Richard-Devantoy et al., 2014) and impaired probabilistic learning in the context of rewards and punishments, including impaired delay discounting (Bridge et al., 2015), impaired reversal learning (Dombrovski et al., 2010) and impaired value comparison during the choice process (Dombrovski et al., 2019); for recent reviews see Lalovic et al. (2022) and Sastre-Buades et al. (2021). Outside of the laboratory, this is corroborated by findings of heightened suicide risk in gambling disorders (Karlsson and Håkansson, 2018; Jolly et al., 2021). Behavioral insensitivity to adverse consequences and heightened sensitivity to internal emotional states have also been linked to suicide attempts (Szanto et al., 2014). Together, these findings have led to a proposal of increased Pavlovian over instrumental control as being an important contributing factor to vulnerability to STB (Dombrovski and Hallquist, 2017, 2021). The Pavlovian controller rigidly specifies stimulus-response mappings regardless of outcomes, such as actively escaping proximal threats and avoiding distal threats, resulting in a rather reflexive behavior. In contrast, the instrumental control specifies stimulus-action-outcome mappings enabling one to adapt behaviors to environmental contingencies and maximize desired outcomes, which can be thought of as goal-directed behavior. In line with the idea of increased Pavlovian biases, a recent study by Millner et al. (2019), found STB to be associated with an increased active-escape bias in an Avoid/Escape Go/No-Go task with aversive sound stimuli. In this study, the STB group was more biased towards choosing an active (Go)

DOI: 10.5334/cpsy.80

response in the presence of an aversive sound (in Escape condition), even when withholding the response (in No-Go condition) was the correct response.

Here, we aim to extend these ideas by proposing a computational mechanism for how the increased Pavlovian biases in STB could result from impaired probabilistic learning (***Figure 1a***). Importantly, we show how this is mediated by hopelessness (a belief that there is nothing one can do to make things better), which is one of the most robust factors of suicide risk (Isometsä, 2014; May et al., 2020). To this end, we apply active inference as the most general neurocomputationally-principled framework that integrates perception, action and learning into a continuous loop of information processing (Friston et al., 2013). The principle guiding this information processing is maximization of (Bayesian) model evidence for one's model of the world, which simultaneously reduces uncertainty about the world and achieves desired outcomes. We show that by operationalizing hopelessness as predominantly negative instrumental beliefs (i.e., with all available actions believed to have low probability of leading to the desired states), an increased Pavlovian control emerges as a straightforward consequence of the drive to maximize model evidence. We propose four different perturbations that within the context of aversive learning could give rise to hopelessness itself: (1) an increased learning from aversive outcomes, (2) a reduced belief decay in response to unexpected outcomes, (3) an increased stress sensitivity and (4) a reduced sense of stressor controllability.

Importantly, these proposals stem from the consideration of neurocircuits implicated in STB. Research on suicide neuromarkers point to the circuits underlying stress response, implicating the locus coeruleus – norepinephrine (LC-NE) and the dorsal raphe nucleus – serotonin (DRN-5-HT) systems (Mann and Rizk, 2020; Oquendo et al., 2014; van Heeringen and Mann, 2014). More broadly, neuroimaging findings are converging on fronto-limbic regions involved in emotion regulation and cognitive control, including the amygdala (Amy), the anterior cingulate cortex (ACC), the dorsal prefrontal cortex (dPFC) and the ventromedial prefrontal cortex (vmPFC) among other regions (Schmaal et al., 2020; Balcioglu and Kose, 2018). However, computational models linking these neuromarkers with the behavioral markers are still missing. Here we suggest that our proposed computational perturbations in STB could be related to how the LC-NE together with the Amy, the dPFC and the ACC mediate learning in response to acute stress and volatility as well as how the DRN-5-HT together with the vmPFC regulate stress responses based on the perceived controllability of the aversive stimulus (***Figure 1b***).

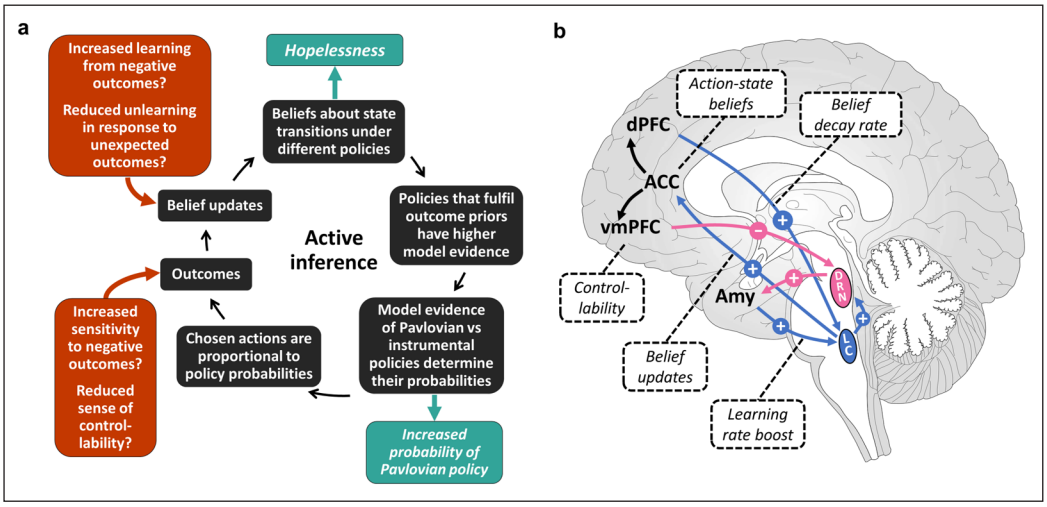

Figure 1 Hypotheses. (**a**) A computational cycle of active inference (black) and potential perturbations at different stages in the cycle (red). These perturbations can give rise to hopelessness – a belief that any taken action will lead to undesired states – and an increased influence of Pavlovian relative to instrumental modes of behavior (teal), both of which are associated with suicidality. (**b**) The brain network that we hypothesize to support the proposed perturbations: norepinephrine modulates belief updates (blue) while serotonin is involved in mediating the effects of stressor controllability (pink). Acute stress leads to increases in the learning rate, which is associated with Amy-LC connectivity (Uematsu et al., 2017; Jacobs et al., 2020), whereas environmental volatility – here assuming state-action prediction errors (SAPEs) as a proxy for environmental change – drives decay of previously learned associations and is mediated by dPFC-LC connectivity (Sales et al., 2019; Clewett et al., 2014). LC projections to the ACC mediate action-dependent state transition belief updates (Tervo et al., 2014; Sales et al., 2019), which are encoded in the ACC (Akam et al., 2021; Holroyd and Yeung, 2012). Finally, controllability of aversive outcomes, which depends on the inferred probabilities of achieving the desired outcomes, reduces aversiveness by inhibiting amygdala activation via the vmPFC-DRN-Amy circuit (Maier and Seligman, 2016; Kerr et al., 2012).

To validate our model, we run model simulations in a probabilistic Avoid/Escape Go/No-Go task, demonstrating how the proposed perturbations lead to hopelessness, increased Pavlovian control and increased active-escape bias – replicating recent empirical findings by Millner et al. (2019). This serves as a proof of concept and produces a computational hypothesis space which can be investigated experimentally and which might speak to different subtypes of suicidal behaviour:

impulsive versus planful attempts (Schmaal et al., 2020; Dombrovski and Hallquist, 2017; Bernanke et al., 2017).

DOI: 10.5334/cpsy.80

2 MATERIALS AND METHODS

2.1 MODELLING RATIONALE

2.1.1 Behavioral control

Within the active inference framework, Pavlovian and instrumental modes of behavior can be derived from the same central computational goal, which could be thought of as maximizing model evidence, resisting entropy or maintaining homeostasis (Pezzulo et al., 2015). Being nested hierarchically – from reflexive to Pavlovian, to habitual, to instrumental behaviors – different modes of behavior allow for the successful navigation of increasingly more complex environments, but also require more computational and metabolic resources. This poses a problem of bounded rationality (i.e. finding a balance between behavioral accuracy and metabolic costs), which can be resolved by performing Bayesian model averaging (BMA) over the different modes of behavior (FitzGerald et al., 2014). This means that actions are informed by all modes of behavior, whereby the modes with the highest model evidence have the most influence. In these computational terms, a stronger active-escape bias in suicidality can be understood as resulting from a reduced model evidence for instrumental relative to Pavlovian control.

In active inference, the model evidence of different policies (e.g., Pavlovian vs. instrumental) depends on how well they are expected to result in desired outcomes (Friston et al., 2016). Thus, saying that instrumental control has a reduced model evidence is the same as saying that instrumental control is expected to have a reduced probability of fulfilling desired outcomes – i.e., beliefs are more 'negative', not mathematically (not below zero), but colloquially speaking. Here we operationalize hopelessness, which is one of the most robust suicide risk factors (May et al., 2020; Isometsä, 2014), as strong negative instrumental beliefs about state transitions.

2.1.2 Learning: uncertainty, stress and norepinephrine

To understand how hopelessness arises, we have to consider the dynamics of belief updating, i.e. learning. Having predominantly negative beliefs (hopelessness) implies either a predominantly aversive environment or preferential learning from aversive events. Asymmetries in how positive and negative outcomes drive learning (i.e. affective bias) have been implicated in mood disorders (Pulcu and Browning, 2017; Clark et al., 2018; Pulcu and Browning, 2019), with negative outcomes having larger effect on learning than positive outcomes (Mathews and MacLeod, 2005; Eshel and Roiser, 2010). Conversely, in the general population learning is driven more strongly by positive outcomes (Sharot and Garrett, 2016). In STB, research on learning from negative vs. positive outcomes is scarce, but a recent study showed STB to be associated with faster processing of negative stimuli (Harfmann et al., 2019).

While the learning rate can be affected by multiple neuromodulatory systems, when it comes to adjusting the learning rate in response to acute stress and volatility, the LC-NE system plays a central role (Pulcu and Browning, 2019; Cook et al., 2019; Silvetti et al., 2018; Jepma et al., 2016; Lawson et al., 2020). Previous influential theories of LC function were founded on the assumption that LC-NE cells behave homogeneously (Yu and Dayan, 2005; Bouret and Sara, 2005). However, recent research emphasizes that LC firing properties are not topographically homogeneous and rather that the LC is comprised of largely non-overlapping target-specific subpopulations of neurons (Poe et al., 2020; Chandler et al., 2019). Importantly, aversive learning is mediated by Amy-LC connectivity (Sterpenich et al., 2006; Uematsu et al., 2017; Jacobs et al., 2020), whereas connectivity between the prefrontal cortex (PFC) regions and the LC has been found to represent 'unlearning', which is necessary for faster adaptation to environmental change or volatility (Uematsu et al., 2017; Sales et al., 2019). Relevant for our aims here, dPFC-LC connectivity has been shown to encode learning from unpredictable feedback (Clewett et al., 2014) and response conflict resolution (Köhler et al., 2016; Grueschow et al., 2020). The dorsolateral PFC (dlPFC) itself has been associated with state prediction error (as opposed to reward prediction error) (Gläscher

DOI: 10.5334/cpsy.80

et al., 2010). LC projections to the ACC have been shown to mediate updates of action-dependent beliefs about the environment (Tervo et al., 2014; Sales et al., 2019), with the ACC encoding such beliefs (Akam et al., 2021; Holroyd and Yeung, 2012). This is consistent with the findings that ACC activity correlates with reward expectation, prediction errors, learning rate and volatility (Rushworth and Behrens, 2008), with these learning variables engaging the ACC primarily in the context of learning about the value of instrumental actions (Matsumoto et al., 2007).

Several lines of evidence suggest the aforementioned networks to be implicated in suicidality (Schmaal et al., 2020; Oquendo et al., 2014). Studies have reported fewer LC neurons, LC overactivity and depletion of NE, all of which are thought to be associated with a dysregulated stress response (Oquendo et al., 2014; van Heeringen and Mann, 2014). The Amy is reported to show increased resting state functional connectivity (Kang et al., 2017) with some structural MRI studies also reporting larger Amy volumes (Monkul et al., 2007; Spoletini et al., 2011). Studies on the dPFC report reduced volumes (Ding et al., 2015), decreased resting regional cerebral blood flow (rCBF) (Willeumier et al., 2011) and reduced activation during error processing (Vanyukov et al., 2016). ACC volumes are also reported to be reduced, with reductions in rostral ACC (rACC) being most significant (Wagner et al., 2011). In a risk aversion task, suicide attempters showed a blunted subgenual ACC activation in response to potential gains (Baek et al., 2017), a reduced ACC response to sad faces and an increased response to wins versus loses (Olié et al., 2015). Finally, a recent study found greater rACC-Amy functional connectivity to be associated with suicidal ideation and previous suicide attempts (Alarcón et al., 2019).

Here, we propose that a disruption in any part of the Amy-dPFC-LC-ACC network (***Figure 1b***, blue) could lead to hopelessness, increased Pavlovian and active-escape bias, increasing the risk of STB. Specifically, we consider two possible perturbations. First, an increased Amy response to negative outcomes would increase learning from negative outcomes (i.e., negative affective bias), which may lead to more negative beliefs (hopelessness) and thus stronger Pavlovian influences. This is supported by increased learning rate in STB observed in an aversive learning task (Millner et al., 2019). Second, reduced activity in the dPFC in response to state-action prediction errors would result in less belief decay allowing negative experiences to accumulate, thus also resulting in hopelessness and stronger Pavlovian biases. Interestingly, impairments in the dPFC have been mostly associated with planful suicides (Schmaal et al., 2020), which would be in agreement with the cognitive rigidity induced by reduced belief decay that we consider here.

2.1.3 Controllability: stress and serotonin

Recent work has shown that controllability of action outcomes governs arbitration between Pavlovian and instrumental control in line with BMA (Dorfman and Gershman, 2019). These effects were found to be associated with frontal midline theta power, which suggests involvement of the mPFC and the ACC (Csifcsák et al., 2020). Furthermore, it has been proposed that dorsal ACC (dACC) could be understood as encoding the expected value of control (Shenhav et al., 2013). This is very similar to what we have proposed in relation to hopelessness. Indeed, *controllability* and *hopelessness* are very closely related constructs. Uncontrollable aversive stimulation has been used to study learned helplessness, from which the construct of hopelessness has been derived (Liu et al., 2015). Another extensively studied effect of controllability is that of modulating the stress response. Stressor controllability has been associated with the vmPFC-DRN-Amy network, and thus with 5-HT-modulated stress response (Maier and Seligman, 2016; Kerr et al., 2012; Hiser and Koenigs, 2018). More specifically, stressor controllability activates the vmPFC, which then inhibits DRN, which in turn reduces amygdala activation in response to a stressor (Maier and Seligman, 2016). Relevant for our aims here, recent studies also show this effect to be associated with successful instrumental learning (Collins et al., 2014; Wanke and Schwabe, 2020).

Considering these findings, we introduce a computational distinction between hopelessness and controllability. As we have defined it earlier, hopelessness corresponds to negative instrumental state-action beliefs that are encoded in the ACC and are arrived at via LC-mediated updates. Controllability, on the other hand, we associate with the vmPFC-DRN-Amy network and thus with 5-HT-modulated stress response. Instrumental state-transition beliefs encoded in the ACC

DOI: 10.5334/cpsy.80

underlie inference about future states and future outcomes, which are then used to estimate controllability in the vmPFC (see Model implementation section for more detailed rationale). This provides a computational link between the NE-modulated and the 5-HT-modulated variables and allows hopelessness and controllability to be distinct but coupled. Interestingly, projections from the LC to the DRN have also been shown to regulate 5-HT release (Pudovkina et al., 2003) and be necessary for developing learned helplessness following uncontrollable stressor exposure (Grahn et al., 2002), providing another point of interaction between the two neuromodulatory systems, which we do not specifically address here.

In suicidality, a large body of research points to deficits in the serotonergic system (van Heeringen and Mann, 2014; Oquendo et al., 2014). While lower 5-hydroxyindoleacetic acid (5-HIAA) levels – a major serotonin metabolite – in the cerebrospinal fluid (CSF) suggest reduced overall serotonergic activity (Mann et al., 2006), serotonin in the brainstem is found to be elevated (Bach et al., 2014), with serotonergic action being elevated in the DRN due to less reuptake (Arango et al., 2001). Furthermore, studies also report elevated serotonin binding in the Amy (Hrdina et al., 1993) and fewer serotonin transporters in the vmPFC and the ACC (Mann et al., 2000). A recent study has also found a history of suicide attempts to be associated with a diminished functional connectivity between vmPFC and Amy (Wang et al., 2020). Together, these findings are consistent with an increased 5-HT-mediated stress response in suicidality.

Here we propose that a reduced sense of controllability stemming from vmPFC-DRN-Amy network impairments (***Figure 1b***, pink) can lead to a stronger Amy activation in response to stress, thus increasing learning from negative outcomes and leading to hopelessness and stronger Pavlovian biases. Impairments in the vmPFC have been associated with impulsive suicide attempts (Schmaal et al., 2020), which would be in line with larger belief updates in response to stressors.

2.2 MODEL IMPLEMENTATION

In the previous sections we have laid out a conceptual picture of our proposed model by considering various computational and neurobiological findings. In this section, we will present one possible computational implementation by focusing on an Avoid/Escape Go/No-Go task (***Figure 2a***). Note that the implementation of the model is not at the level of neural dynamics but rather at the higher level of computational mechanisms underwritten by such dynamics (cf. Marr's levels of analysis (Marr and Poggio, 1976)). However, the active inference framework has deep connections to neurobiology and has recently been applied to understanding a whole range of psychiatric conditions (Smith et al., 2021), including the effects of noradrenergic and serotonergic drugs in depression (Constant et al., 2021).

The task employs a 2 × 2 (Go/No-go x Avoid/Escape) factorial design. On every trial, the agent is presented with one of four cues. Two of the cues are always paired with an aversive sound (Escape condition) while the other two are paired with silence (Avoid condition). The agent's goal is to learn, for each cue, which response (active Go or passive No-go) more frequently results in silence during feedback. For ease of reference, we will refer to the responses that maximize the frequency of silence during the feedback as the "correct" responses throughout the paper. This means that in the Avoid condition, correct responses will prevent the aversive sound from playing, while in the the Escape condition, correct responses will stop the aversive sound that is already being played. However, the feedback is probabilistic, which means that even "correct" responses will sometimes lead to experiencing the aversive sound. Probabilistic feedback introduces uncertainty and makes it more challenging to learn, which response is correct.

To model this task, we use an active inference scheme for discrete Markovian models (Friston et al., 2016), which means that we will be dealing with *discrete* time steps (t), states (s), actions (a), and observations (o). Each trial gets divided into three time steps. At $t = 1$, the agent is in one of four possible hidden states with no informative observations about the task conditions available (e.g., a fixation cross is displayed, which does not contain any information on which cue will be presented next). At $t = 2$, the agent is presented with one of the four cues, which correspond to one of the four conditions resulting from the 2 × 2 (Go/No-Go × Avoid/Escape) factorial design. In the Avoid

DOI: 10.5334/cpsy.80

condition, there is no aversive sound while the agent is choosing either a Go or No-Go response. In the Escape condition, the aversive sound is present throughout the decision phase. At $t = 3$, the agent observes the final outcome of a trial, either aversive or neutral. This means that in the Avoid condition, a correct action leads to no aversive sound, while in the Escape condition, a correct action results in the discontinuation of the aversive sound. When choosing an action at $t = 2$, the agent relies on available policies π: instrumental Go/No-Go and Pavlovian. Probabilities of these policies depend on the underlying beliefs about likelihood of observations, **A**, state transitions – **B**{Go}, **B**{No-Go}, $\mathbf{B}_0$ – as well as prior beliefs over outcomes (i.e., preferences), **C**. In other words, probabilities of policies depend on model evidence that each set of beliefs provides, where model evidence is approximated with variational free energy (see **S1 Appendix: full mathematical details of the model** and Da Costa et al. (2020) for more details).

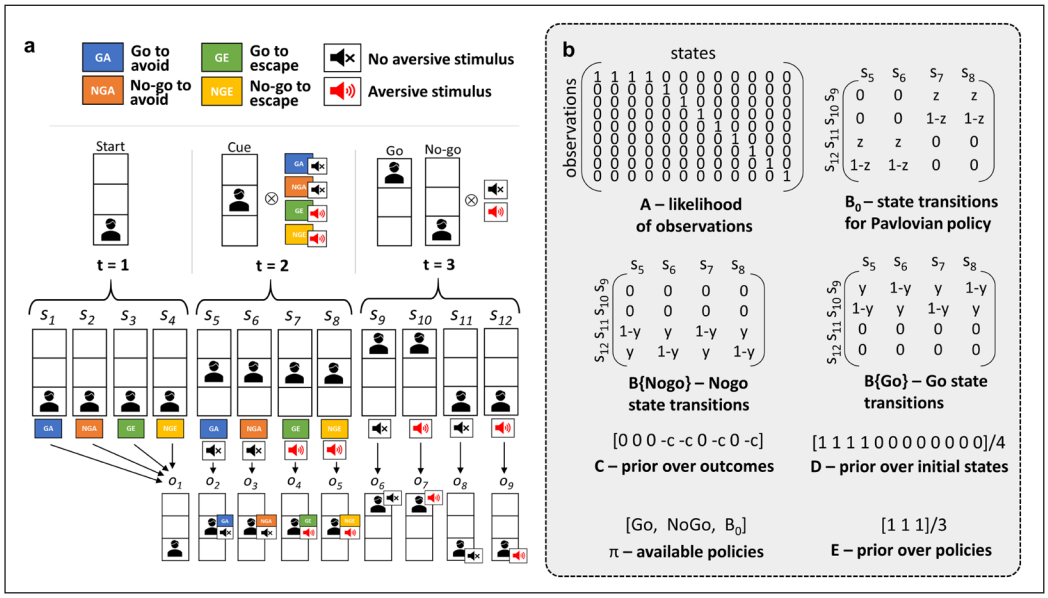

Figure 2 Avoid/Escape Go/No-Go task design and model specification. (**a**) Following Millner et al. (2018; 2019) the task has 4 cues corresponding to the 2 × 2 (Go/No-Go x Avoid/Escape) factorial task structure, with 2 possible outcomes: aversive or neutral. For modelling purposes, the task was divided into 3 discrete time points. At the start of a trial ($t = 1$) the agent is in one of the four hidden states (s_{1-4}) and no observations are available (o_1). Next, the agent is taken to $t = 2$, where a cue (and in the case of Escape condition also an aversive sound) is presented corresponding to one of four possible hidden states (s_{5-8}) and observations (o_{2-5}). At $t = 2$ the agent chooses what action to take (Go or No-Go) which then leads to one of four possible states (s_{9-12}) and observations (o_{6-9}): Go response + silence (s_9, o_6), Go response + aversive sound (s_{10}, o_7), No-Go response + silence (s_{11}, o_8), No-Go response + aversive sound (s_{12}, o_9). (**b**) The main model structures. The likelihood of observations, **A**, was implemented to have deterministic mappings between states and observations due to the salience of the aversive stimulus and the cues. As no learning was required, **A** in the generative model and in the generative process were identical. State transitions from $t = 2$ to $t = 3$ for instrumental (Go/No-Go) policies B were probabilistic, captured by the y parameter. For the objective transition probabilities y was set to 0.8, meaning that correct response by the agent led to the neutral state 80% of the time. For the generative model, y was initialized with 0.5 to correspond to the agent having a uniform prior over the two possible transitions. The zero probabilities for the other transitions reflect the assumption that the agent understands the task structure and does not expect to end up in a Go state after choosing No-Go and vice versa. State transition probabilities from $t = 2$ to $t = 3$ for the Pavlovian policy, $\mathbf{B}_0$, were implemented to allow only for No-Go responses in the Avoid and Go responses in the Escape conditions (with Go responses in the Avoid and No-Go responses in the Escape conditions having 0 probabilities). The strength of the belief that the Pavlovian policy will lead to the desired states is captured by the z parameter. Prior over outcomes (**C**) assumed that the agent does not like outcomes 4, 5, 7 and 9 (all of which involve the aversive stimulus). The strength of this preference of neutral outcomes is captured by parameter c. The prior over initial states **D** was assumed to be uniform for states 1–4. The other states have zero probability, which reflects the assumption that the agent understands the task structure and does not expect to be in states 5–12 at the beginning of a trial. Finally, prior over policies **E** was also assumed to be uniform across the available Go, No-Go and Pavlovian policies (π). See **S1 Appendix: full mathematical details of the model** for more implementation details.

DOI: 10.5334/cpsy.80

After each trial, the agent updates their beliefs depending on the outcome in that trial. Since there is no ambiguity about observations due to their saliency, we assume all learning to concern only state transition probabilities (**B**). Columns in **B** matrices are Dirichlet distributions parameterized with concentration parameters $\boldsymbol{b}$, such that for a control state u, $\boldsymbol{B}(u) = Dir(\boldsymbol{b}(u))$. Concentration parameters can be interpreted as the number of times various combinations of state transitions have been observed, which effectively captures both the probability and the confidence in that probability. At the end of each trial, state transition concentration parameters are updated via:

$$\boldsymbol{b}_i(u) = b_{i-1}(u) + \eta \sum_{\tau,p} \boldsymbol{\pi}^{\boldsymbol{p}}_{\tau-1}\, \boldsymbol{s}^{\boldsymbol{p}}_{\tau} \otimes \boldsymbol{s}^{\boldsymbol{p}}_{\tau-1} - \frac{b_{i-1}(u)-1}{\lambda}, \tag{1}$$

where i denotes the trial number, u denotes the control state (Go or No-Go) and $\boldsymbol{s}^{\boldsymbol{p}}_{\tau}$ contains posterior probabilities of different states under each policy p for time point τ. Note that in the current implementation, we only care about $\tau = 3$, because the transition between $t = 1$ and $t = 2$ does not depend on the agent's choices. π denotes posterior policy probabilities.

The sum in the second term of the equation is performed only over the two instrumental policies. This means that in an extreme case where behavior is driven primarily by the Pavlovian policy and the probabilities of instrumental policies are very low, there will be very little learning even though the agent has the information to update their beliefs about state transitions. To account for instrumental learning facilitated by Pavlovian responses (Holmes et al., 2010), one could consider combining the posterior probabilities of Pavlovian Go or No-Go responses with instrumental Go and No-Go policy probabilities, respectively, when updating beliefs about controlled state transitions. However, in the simulations presented in this paper, Pavlovian effects are never too extreme and similar results can be obtained with either implementation. To keep the model simpler, here we present the results using the original implementation where the sum in the second term of the equation is performed only over the two instrumental policies.

The remaining two parameters η and λ in **Eq. (1)** control the learning rate and the decay rate, respectively. The learning rate controls how much new experiences add to the existing concentration parameters, while the decay rate controls how much the previously accumulated concentration parameters should be discounted. Without the decay factor, concentration parameters would accumulate indefinitely making the agent too rigid and thus too slow to adapt if environmental contingencies were to change. Following the work of Sales et al. (2019), λ is assumed to depend on state-action prediction errors (SAPEs) and to be associated with effective connectivity from the dPFC to the LC. This makes the decay factor sensitive to environmental volatility: changing environmental contingencies will result in larger SAPEs, which in turn will speed up unlearning of no longer accurate beliefs, allowing the agent to learn the new contingencies faster. The relationship between SAPEs and λ is modelled using a logistic function:

$$\lambda = \lambda_{min} + \frac{\lambda_{max} - \lambda_{min}}{1 + e^{\,g(SAPE - m)}}, \tag{2}$$

where g is the gradient, m is the midpoint, while λ_{min} and λ_{max} are minimum and maximum function values. Note that higher SAPEs will result in a smaller λ, which will result in more belief decay because λ is a denominator in the update equation **Eq. (1)**. SAPE itself is defined as Kullback-Leibler (KL) divergence between BMA distributions at successive time steps:

$$SAPE(t) = D_{KL}[(\boldsymbol{S}^{\boldsymbol{t}}_{\tau}) \,||\, \boldsymbol{S}^{\boldsymbol{t-1}}_{\tau})] \ . \tag{3}$$

In the simulations presented in this paper, SAPE is computed for t = 3, after the action (Go/No-Go) is performed and only for predictions about the final states ($\tau = 3$). BMAs themselves are computed via:

$$\boldsymbol{S}_{\tau} = \sum_{p} \boldsymbol{\pi}^{\boldsymbol{p}}_{\tau} \cdot \boldsymbol{s}^{\boldsymbol{p}}_{\tau}, \tag{4}$$

where $\boldsymbol{\pi}^{\boldsymbol{p}}_{\tau}$ denotes posterior policy probabilities and $\boldsymbol{s}^{\boldsymbol{p}}_{\tau}$ denotes posterior state probabilities for policy p at time point τ.

DOI: 10.5334/cpsy.80

In addition to being sensitive to environmental change (i.e. volatility), the LC-NE system also coordinates aversive learning mediated by Amy-LC connectivity (Uematsu et al., 2017; Jacobs et al., 2020). To capture these effects, we introduce a learning rate dependency on outcome valence (assuming Amy activation during aversive outcomes), which we associate with the preference against aversive outcomes encoded in the **C** vector:

$$\eta = 1 + k\,|\,\boldsymbol{C}(o)\,|\,, \tag{5}$$

where *C(o)* is the value of prior preference for outcome *o*, with the parameterization being *-c* for the aversive stimulus outcomes and *0* for the neutral outcomes. Parameter *k* is a scaling factor that could correspond to effective connectivity between the Amy and the LC. Note that the learning rate dependence on valence that we introduce here is what enables the model to account for affective biases (Pulcu and Browning, 2017; 2019; Sharot and Garrett, 2016; Eshel and Roiser, 2010). A more principled implementation of valence and its role in modulating the learning rate could depend on the rate of change of free energy over time (Joffily and Coricelli, 2013).

The final component of the model aims to account for how controllability of aversive outcomes inhibits Amy activation via the serotonergic system involving vmPFC-DRN-Amy network (Maier and Seligman, 2016; Kerr et al., 2012). We implement this by modulating stress sensitivity parameter *c* by a controllability parameter *w*:

$$c' = c^{(1-w)} \tag{6}$$

In the limiting cases when there is no control ($w = 0$), c' is equal to the original *c* and when there is complete control ($w = 1$) c' is equal to 1. Controllability itself is assumed to depend on on beliefs that the neutral outcome can be reached:

$$w_n = \sum_{i=6,8} \boldsymbol{o}_{\tau=3}^{t=2}(o_i)\,, \tag{7}$$

where $\boldsymbol{o}_{\tau=3}^{t=2}$ contains expected outcomes at time point $\tau = 3$ at time $t = 2$; and here we are summing over the two possible neutral outcomes (o_6 and o_8) for time point $\tau = 3$. Expected outcomes are simply a product of the likelihood of observations **A** and BMA expected states $\boldsymbol{S}\tau$. This means that the subjective estimate of controllability depends on inferred and expected states, and not on actual states of the world. Parameter w_n effectively represents an average probability of achieving the desired outcome in the inferred context. Note that this is similar to the well-established finding of vmPFC encoding expected value (see Hiser and Koenigs (2018) for a review). Furthermore, such distinction between vmPFC, which encodes expected outcome (which we associate with controllability), and ACC, which encodes state-transition probabilities (which we relate to hopelessness) is consistent with the finding that vmPFC encodes stimulus-based value and is more active during the outcome phase (cf. stress response) and that ACC encodes action-based value and is more active during both outcome and decision phases (cf. instrumental control and learning) (Vassena et al., 2014). The close relationship between the subjective feeling of control and outcome valuation has also been demonstrated in recent studies (Stolz et al., 2020; Wang and Delgado, 2019). Relevantly, STB has been associated with reduced activation to expected value in vmPFC (Brown et al., 2020; Dombrovski and Hallquist, 2017).

Finally, to collectively account for any impairments of how w_n modulates the stress response (i.e., any impairments along the vmPFC-DRN-Amy network), we transform w_n into the final estimate of controllability by entering it into a logistic function constrained by a controllability threshold w_0 (i.e. the midpoint of the logistic function) and a gradient g_w:

$$w = \frac{1}{1 + e^{\,-g_w(w_n - w_0)}}\,. \tag{8}$$

The dependency of the learning rate on stress sensitivity (*c*), means that controllability can indirectly regulate learning rate through its effects on stress sensitivity. This is in line with recent findings

DOI: 10.5334/cpsy.80

showing that DRN serotonin neurons modulate the learning rate and they do so in proportion to uncertainty about decision outcomes (Grossman et al., 2021), which is also in agreement with how we implemented controllability (**Eq. (7)**).

Note that even though we have provided a reasonable theoretical justification for introducing the controllability component, it is a rather ad hoc addition to the otherwise computationally principled active inference framework. It is important to stress, however, that the simulation results that we present in the next section do not hinge on this additional computation, except for the results concerning the controllability parameter itself.

Figure 3 summarizes the proposed computations as well as their possible neural correlates and highlights parameters of interest for STB.

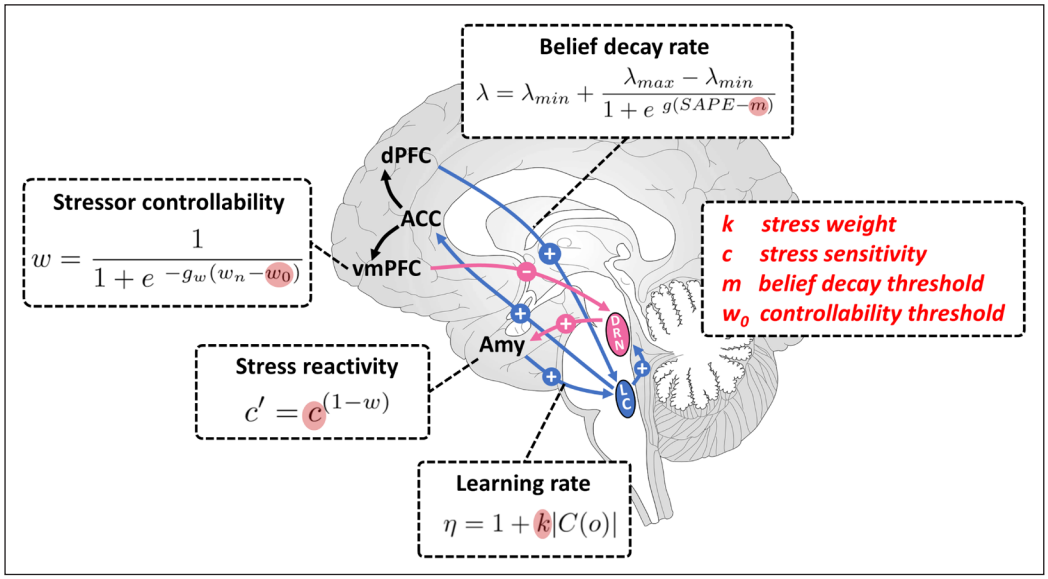

Figure 3 Summary of the proposed computations, possible neural correlates and parameters of interest for STB. Within the proposed model there are four areas of relevance for STB: learning rate, belief decay rate, stress reactivity and perceived controllability of a stressor. Stress weight parameter, *k*, controls the boost in the learning rate in response to stress. Increasing this parameter would result in increased learning from stressful outcomes. Stress sensitivity parameter, *c*, captures individual sensitivity to stress, which then also affects the learning rate. Controllability threshold, w_0, is a midpoint in the logistic function that translates the beliefs about state transitions into an estimate of stressor controllability. In other words, w_0 regulates how positive state transition beliefs have to be for a stressor to be deemed sufficiently controllable. Finally, belief decay threshold, *m*, regulates how large state-action prediction errors (SAPEs) have to be before significant belief decay takes place. Note that for the decay rate and the controllability there are other parameters (gradients, g_w, *g*, and minimum and maximum decay values λ_{min}, λ_{max}) that we could inspect, but for simplicity here we focus on the midpoint values w_0 and *m* as the exact parameterization of these effects is somewhat arbitrary and the midpoints are sufficient for exploring the general direction of different manipulations.

3 RESULTS

3.1 MODEL SIMULATIONS

To validate the model, we first simulated performance on the task for a single healthy control (***Figure 4***), and then showed how increasing parameter *k* – which regulates aversiveness-related component of the learning rate and is assumably represented in terms of Amy-LC connectivity – can produce increased active-escape biases and other behavioral and cognitive aspects associated with suicidality (***Figure 5***). Finally, we defined a wider hypothesis space, exploring how different parameters in the model can independently lead to the behavior observed in STB. (***Figure 6***).

For the initial simulations (***Figures 4*** and ***5***), we ran 200 trials of the task, where at every trial one of the 4 cues was presented at random. After 100 trials, the meanings of the cues were reversed: Go becoming No-Go and vice versa. In this simulation, the model parameters were set to $k = 0.1$, $m = 1.3$, $c = 8$, $w_0 = 0.5$, $z = 0.4$, $\lambda_{min} = 2$, $\lambda_{max} = 50$, $\alpha = 3$ and $\beta = 1$ to produce reasonable performance trajectories as well as an active-escape bias (***Figure 4a***) consistent with empirical findings reported by Millner et al. (2018). As the agent's beliefs approach the actual state transition probabilities (***Figure 4f-i***, colored lines), this makes the neutral outcomes more expected, thus invoking only small SAPEs in contrast to unexpected aversive outcomes (***Figure 4d***). This is also what drives successful unlearning after the reversal: a series of negative outcomes with large SAPEs result in a sharp drop in the decay parameter (***Figure 4e***, black line), which increases belief decay and facilitates quick learning of new contingencies. The Pavlovian policy that underlies the active-escape bias can be seen at its strongest at the very beginning of the task and right after the reversal, when beliefs that instrumental actions will lead to neutral outcomes are lower (***Figure 4f-i***).

DOI: 10.5334/cpsy.80

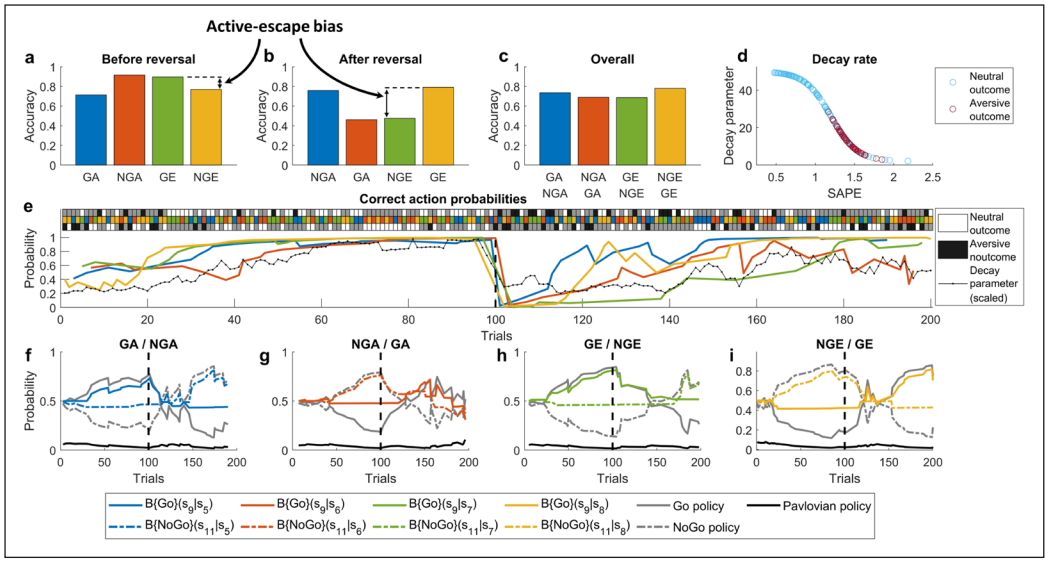

Figure 4 Model simulations: a single (healthy control) participant with low stress weight (k = 0.1). (**a-c**) average choice accuracy before reversal, after reversal and overall, respectively, for Go-to-Avoid (GA), No-Go-to-Avoid (NGA), Go-to-Escape (GE) and No-Go-to-Escape (NGE); the four colors denote different cues used in the task. The results in (**a**) reproduce active-escape bias reported by Millner et al. (2018) in the general population. (**b**) and (**c**) are additional predictions about performance after reversal and overall, respectively. (**d**) Decay parameter values for different SAPEs throughout the task. Note that SAPEs for aversive outcomes are larger which leads to smaller decay parameter, and thus to larger belief decay (see **Eq. (1)**). (**e**) Performance across all trials. The top 3-row panel shows the sequence of cue presentation (middle row), executed action (non-grey squares: bottom row – No-Go, top row – Go) and trial outcome (white – neutral, black – aversive); each column corresponds to a single trial. Actions are represented implicitly by either black or white color. If for a given trial the top square is either black or white, it means that the Go action was selected, if the bottom square is either black or white then the No-Go action was selected. The main panel shows trajectories of correct action probabilities, which gradually increase as the task progresses, but drop sharply once the Go/No-Go cue meanings are reversed on trial 100. The response to this environmental change can be seen in the decreased decay parameter (black line), which drives faster forgetting of previously learned contingencies and allows the agent to adapt. Note that decay parameter trajectory here is scaled to be between 0 and 1 and smoothed out using moving average with a window size of 5 trials. (**f-i**) Trajectories of underlying beliefs about state transitions and policy probabilities. These plots reflect the straightforward relationship between belief strength and policy probability: as the probability of an instrumental Go/No-Go action leading to the desired state increases (solid/dash-dotted colored lines) the probability of choosing Go/No-Go policy tracks that increase (solid/dash-dotted gray), and probabilities of Pavlovian policies (solid black) decrease as a result. The vertical dashed lines in all of the plots denote the reversal.

By increasing parameter k to 1, the size of the belief update after experiencing aversive outcomes becomes larger, reproducing the increased active-escape bias (***Figure 5a***) by a similar magnitude as reported in individuals with STB (Millner et al., 2019). The increase in the active-escape bias is a direct consequence of the increased influence of the Pavlovian policy (***Figure 5f-i***, black line), which in turn is a consequence of weaker beliefs that either of the instrumental Go/No-Go actions will lead to the desired neutral outcome (cf. hopelessness) (***Figure 5f-i***, colored lines). The latter is a direct consequence of increased k, leading to an over-adjustment of beliefs after aversive outcomes. This also disrupts the agent's ability to adapt to a changing environment because negative outcomes after the reversal become less surprising: this is reflected in reduced SAPEs for aversive outcomes and increased SAPEs for neutral outcomes (***Figure 5d***). Assuming SAPEs are computed in dPFC (Sales et al., 2019; Gläscher et al., 2010), this result would be consistent with empirical findings of increased dPFC response to wins vs. loses in suicide attempters (Olié et al., 2015) and reduced dlPFC activation in response to negative stimuli in suicidal ideation (Miller et al., 2018).

DOI: 10.5334/cpsy.80

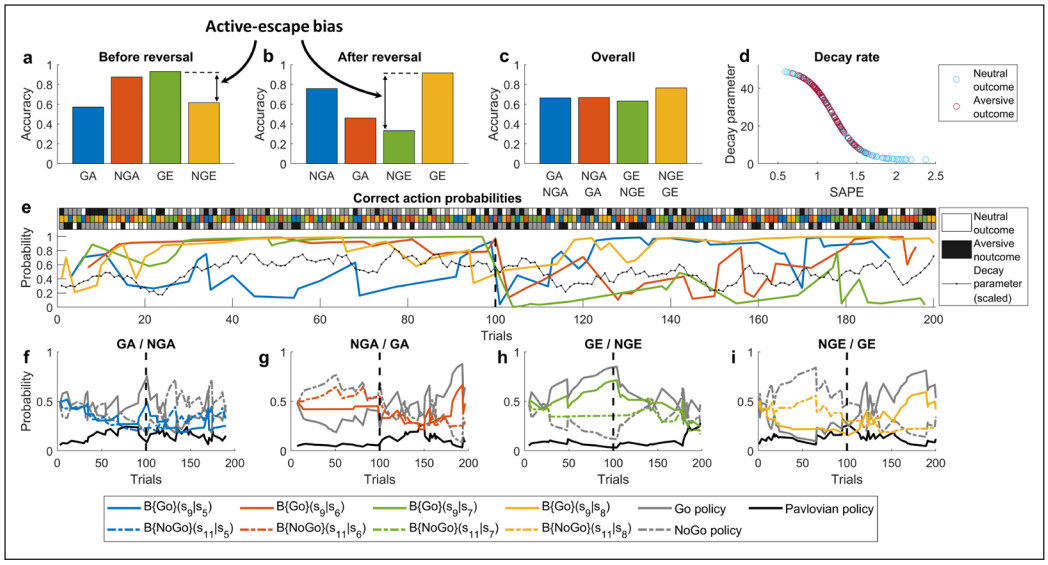

Figure 5 Model simulations: a single (STB) participant with a high stress weight (*k = 1*). (**a-c**) average choice accuracy before reversal, after reversal and overall, respectively, for Go-to-Avoid (GA), No-Go-to-Avoid (NGA), Go-to-Escape (GE) and No-Go-to-Escape (NGE); the four colors denote different cues used in the task. The results in (**a**) reproduce increased active-escape bias in suicidality reported by Millner et al. (2019), and predict that this bias would be even larger after a reversal in cue meanings (**b** panel). (**d**) Decay parameter values for different SAPEs throughout the task. Note that now aversive outcomes produce smaller SAPEs, due to increased expectation of aversive states. (**e**) Performance across all trials. The top 3-row panel shows the sequence of cue presentation (middle row), executed action (non-grey squares: bottom row – No-Go, top row – Go) and trial outcome (white – neutral, black – aversive); each column corresponds to a single trial. Actions are represented implicitly by either black or white color. If for a given trial the top square is either black or white, it means that the Go action was selected, if the bottom square is either black or white then the No-Go action was selected. The main panel shows trajectories of correct action probabilities. Compared to the healthy control in the previous figure, the trajectories are noisier, especially after the reversal on trial 100. Decay rate trajectory (black line) is also nosier, which is partly responsible for the poor adaptation after the reversal. Note that decay parameter trajectory here is scaled to be between 0 and 1 and smoothed out using moving average with a window size of 5 trials. (**f-i**) Trajectories of underlying beliefs about state transitions and policy probabilities. Compared to the healthy control, the belief trajectories are noisier, but even more importantly, beliefs about the instrumental transitions to neutral states are on average weaker (cf. hopelessness), which leads to increased probability of the Pavlovian policy. The vertical dashed lines in all of the plots denote the reversal.

3.2 MULTIPLE ROUTES TO AN ACTIVE-ESCAPE BIAS

While directly increasing learning from aversive outcomes (*k*) is one way to produce the effects associated with STB, there is a wider hypothesis space to be explored. To that end, we performed a more extensive investigation of the effects of other model parameters. In this context, it is important to note that the model exhibits a considerable degree of stochasticity when initiated with the chosen parameter configurations and thus, the results presented earlier in ***Figures 4*** and ***5*** are meant to be primarily illustrative. To reduce stochasticity and to obtain more robust behavioral results, now we used 400 trials with a reversal at 200 and ran 50 simulations for each parameter configuration. To visualize the results, we computed relevant task performance summary statistics (mean and standard error) for each parameter configuration (***Figure 6***). The first column in ***Figure 6*** simply reproduces the results in ***Figures 4*** and ***5***, showing that as we increase learning from negative outcomes, we reduce beliefs that instrumental actions will lead to the desired states (***Figure 6a***), which leads to an increase in the probability of the Pavlovian policy (***Figure 6e***), which in turn leads to a larger active-escape bias (***Figure 6i***). As a result of the increased biases, we also see a slight decrease in the overall performance accuracy (***Figure 6m***).

Reducing base belief decay (increasing parameter *m*) produces similar results of more negative beliefs, a higher probability of the Pavlovian policy and a stronger active-escape bias (***Figure 6b,f,j***). We also see a deterioration of the overall performance accuracy after the reversal, as the agent

is slow to adapt to new contingencies (***Figure 6n***). Although very little research exists on reversal learning in suicidality, the latter result is in line with impaired reversal learning demonstrated in a reward/punishment probabilistic learning task in suicide attempters (Dombrovski et al., 2010).

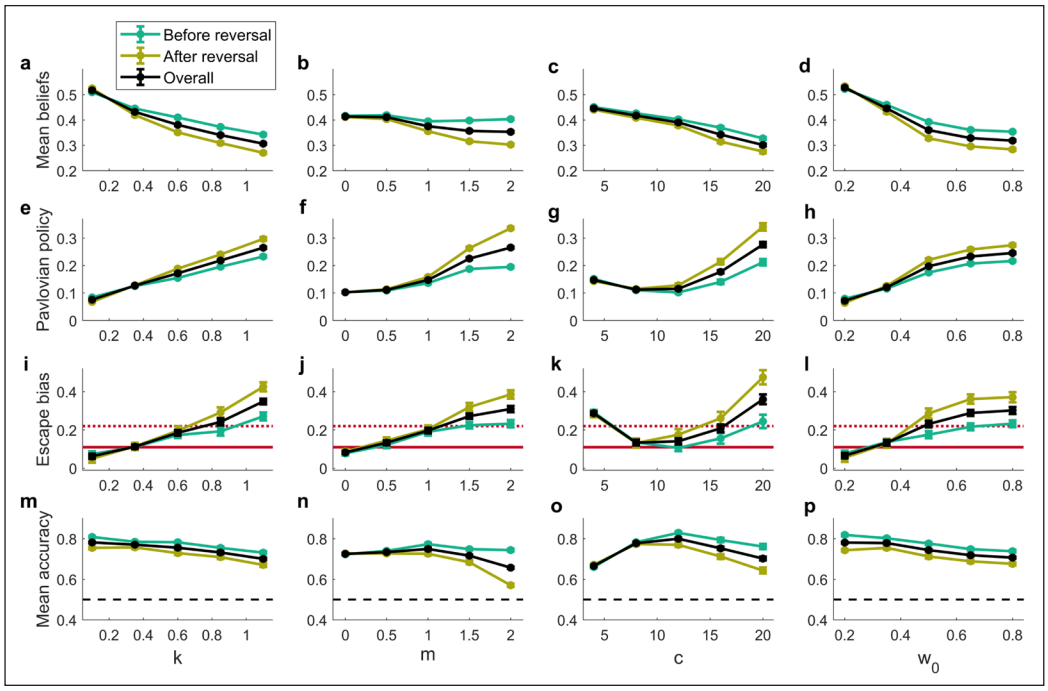

Figure 6 Model simulations: exploration of the hypothesis space. Each column shows the effects of varying on of the parameters: k – stress weight (while $m = 1.3$, $c = 8$, $w_0 = 0.6$), m – belief decay threshold (while $k = 0.7$, $c = 8$, $w_0 = 0.6$), c – stress sensitivity (while $k = 0.6$ $m = 1$, $w_0 = 0.5$) and w_0 – controllability threshold (while $k = 0.9$, $m = 1.3$, $c = 8$). (a-d) the mean of beliefs that the neutral state will be reached averaged across 4 contexts and 2 possible actions. (**e-h**) The mean probability of choosing the Pavlovian policy. (**i-l**) Active-escape bias (the difference between choice accuracy on GE and NGE trials). The solid and dashed red lines denote the expected active-escape bias in healthy control group and suicidality group, respectively (based on Millner et al. (2018; 2019) findings). (**m-p**) Mean choice accuracy across all 4 contexts.

A higher stress sensitivity (larger c) also produces the effects associated with STB: more negative beliefs (***Figure 6c***) lead to a higher probability of the Pavlovian policy (***Figure 6g***) and a stronger active-escape bias (***Figure 6k***). Finally, the overall performance accuracy (***Figure 6o***) shows a non-linear dependence on stress sensitivity, which is reminiscent of the inverted U-shaped relationship between stress and performance (Yerkes et al., 1908; Hebb, 1955). It is important to note that the c parameter features in the model twice: first, in the prior over outcomes, and second, in the learning rate after aversive outcomes. The decrease in the overall performance accuracy and the increase in the active-escape bias at very low values of c can be explained by the former role of this parameter. In other words, a small c means little motivation to prefer neutral outcomes (e.g., the aversive outcomes are not experienced as very aversive), which leads to a more random policy selection and thus effectively increases Pavlovian influences and reduces overall performance accuracy. In contrast, the increased active-escape bias associated with larger c values derives from parameter c's contribution to the learning rate. Interestingly, both reduced and increased distress tolerance have been associated with suicide risk: lower distress tolerance relates to psychological/social pain and contributes to suicidal ideation, while higher distress tolerance relates to physical pain and contributes to the acquired capability for suicide (see Liu et al. (2016) for a discussion). Our model simulations are agnostic to the nature of the aversive stimulus used and thus might be capturing both of these effects.

Reducing perceived controllability (increasing w_0) is yet another way to produce the effects associated with STB. By way of a self-fulfilling prophecy, a reduced controllability threshold leads to more negative beliefs (***Figure 6d***), which induces increases in the Pavlovian policy probability (***Figure 6h***) and an active-escape bias (***Figure 6l***), as well as a slight decrease in the overall performance accuracy (***Figure 6p***).

3.3 COMPUTATIONAL PARAMETERS REVEAL STB SUBTYPES

While all of the above parameter manipulations lead to similar mean behavioral effects, inspecting the time series reveals different dynamics of belief updating and policy probabilities (***Figure 7***). Using NGE/NE cue as an example, for high m values (low belief decay rate), we can see a very

gradual progression towards more negative beliefs and an increased influence of the Pavlovian policy (***Figure 7a***). For high w_0 (low controllability), high k (high stress weight) and high c (high stress sensitivity), we see increasingly larger and sudden spikes in Pavlovian biases. Considering the influence of the Pavlovian policy as a proxy for STB risk, the former scenario suggests a constantly increasing risk of STB and thus could be related to planful suicide attempts, while the latter scenario suggests an increased STB risk immediately after the occurrence of aversive events and could relate to impulsive suicide attempts. Bearing in mind our proposed links between the model parameters and the underlying neurocircuitry (***Figure 3***), these results are consistent with planful and impulsive suicide attempt subtypes, with the former being predominantly associated with dPFC activity and the latter being predominantly associated with vmPFC activity (Schmaal et al., 2020).

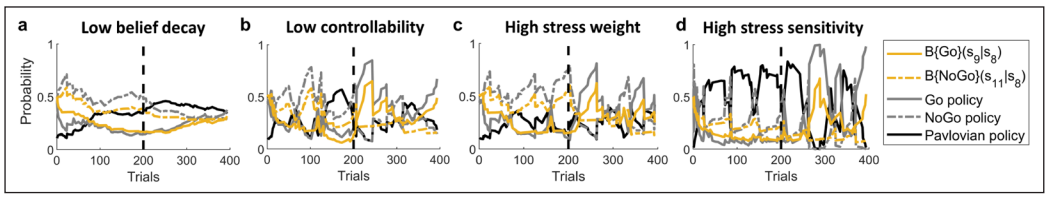

Figure 7 Model simulations: trajectories of beliefs and policies under different parameter manipulations. (**a**) low belief decay, $m = 2$, (**b**) low controllability, $w_0 = 0.8$, (**c**) high stress weight, $k = 1.1$, (**d**) high stress sensitivity, $c = 20$. The other parameters were set to the same values as in Figure 6. All panels show trajectories of NGE/NE cue: where the cue is NGE before the reversal (the vertical dashed line) and GE after the reversal. Less variable rigid negative beliefs and Pavlovian policy in (a) could be associated with planful suicide attempts, whereas more variable beliefs and sudden increases in Pavlovian policy in (**b-d**) could be associated with more impulsive suicide attempts (Schmaal et al., 2020; Bernanke et al., 2017).

4 DISCUSSION

In this paper, we presented a computational model of hopelessness and Pavlovian/active-escape bias in suicidality. With this model we showed that increased Pavlovian control and active-escape biases result from state hopelessness via the drive to maximize model evidence. Moreover, we proposed how hopelessness itself can arise from four mechanisms: (1) increased learning from aversive outcomes, (2) reduced belief decay in response to unexpected outcomes, (3) increased stress sensitivity and (4) reduced sense of stressor controllability. We also proposed how these alterations might relate to the neurocircuits implicated in suicidality. Specifically, we considered perturbations in the LC-NE system together with the Amy, the dPFC and the ACC, which mediate learning in response to acute stress and volatility, as well as perturbations in the DRN-5-HT system together with the vmPFC and the Amy, which regulate stress reactivity and its modulation by perceived controllability. We validated the model via simulations of an Avoid/Escape Go/No-Go task reproducing the active-escape biases reported by Millner and colleagues (Millner et al., 2019, 2018).

First, it is worthwhile to elaborate on what advantages and new insights our proposed model brings compared to previous modelling work. Millner et al. (2019) analyzed the increased active-escape bias in STB using a combined reinforcement learning – drift diffusion model (RL-DDM) and found that an increased active-escape bias can be explained by a bias parameter (aka a starting point in the DDM part of the model). This parameter was assumed to be constant throughout the task. In contrast, our proposed model offers a mechanistic explanation for how active-escape bias arises dynamically from learning about the state transition probabilities and balancing between instrumental and Pavlovian policies. Unlike in RL-DDM, in our model, Pavlovian and instrumental policies are represented explicitly. Importantly, this allowed us to relate state transition probabilities to state hopelessness (which is a central construct in suicidality research (Klonsky et al., 2018; May et al., 2020; Isometsä, 2014)), offering a possible operationalization of the hopelessness construct. Finally, using the active inference framework enabled us to propose several links (some more speculative than others) between the model variables and the underlying neurocircuitry, which could help bridge the explanatory gap between neurobiology and cognition in STB (see Limitations section for further discussion).

Our model simulation results offer a computational hypothesis space by identifying mechanistically distinct perturbations that lead to hopelessness and Pavlovian/active-escape biases associated with STB. These distinct pathways might also speak to different suicidality subtypes: impulsive versus planful (Schmaal et al., 2020; Bernanke et al., 2017). While all of the four parameter manipulations produced increased Pavlovian control and active-escape biases, examining the trajectories of belief updating revealed that reduced belief decay led to more gradual updates and more stable negative beliefs as well as more stable and elevated Pavlovian influences, which could be associated with

DOI: 10.5334/cpsy.80

more planful STB. The other three manipulations – reduced controllability of stressors, increased learning from aversive outcomes and increased stress sensitivity – resulted in increasingly variable belief updates with sudden spikes in Pavlovian biases after aversive outcomes, which could be associated with more impulsive STB. Considering the dPFC and the vmPFC as possible correlates of belief decay and controllability (and its effects on stress reactivity), respectively, our results are in agreement with neuroimaging studies associating disruptions in vmPFC activity with the impulsive STB subtype and the dPFC activity with the planful STB subtype (Schmaal et al., 2020).

While throughout the paper we have adopted a transdiagnostic view of STB, many mental disorders are known to increase suicide risk. Among all disorders, borderline personality disorder (BPD), depression, bipolar disorder, schizophrenia, and anorexia nervosa show the highest risk of suicide – between 10 to 45 times higher than the general population (Chesney et al., 2014). Comorbidities further increase suicide risk by inflicting higher levels of distress (Nock et al., 2010; Jylhä et al., 2016), with the majority of suicides being estimated to occur within a major depressive episode (Isometsä, 2014). Recent studies show preliminary evidence that suicide subtypes might cut across the current categories of disorders, with higher suicidal ideation variability (i.e. higher stress responsiveness) being associated with childhood physical abuse, aggression, and impulsivity in major depressive disorder (Oquendo et al., 2020) and with affective lability in BDP (Rizk et al., 2019). In a similar way, we might expect that the ways in which different mental disorders increase the risk of suicide could also map onto the different ways in which the effects associated with STB can emerge within our proposed model.

4.1 MODEL-BASED SUICIDALITY SUBTYPES AND PERSONALIZED INTERVENTIONS

Being able to stratify the propensity for suicidal behavior into mechanistically distinct subgroups could help improve early interventions and treatment response prediction. Many different psychotherapies are applied in the context of suicidality, including the manualized therapies such as CBT, Dialectical Behavior Therapy (DTB), and mentalization-based therapy (MTB). However, evidence for the effectiveness of different psychotherapies is still scarce and it remains unclear which components of the therapies are most effective in reducing suicidality (Briggs et al., 2019; Ougrin et al., 2015; Weinberg et al., 2010). Moreover, the attempts to determine these unknowns are likely complicated by not accounting for the etiological heterogeneity in high suicide risk groups (Iyengar et al., 2018). Current neurobiological models of the mechanism of action of psychotherapy point to neural substrates of executive and semantic processes and highlight the vmPFC and its involvement in implicit emotion regulation as well as dPFC and its involvement in explicit behavioral control (Messina et al., 2016). This would map to the stressor controllability (vmPFC) and belief decay (dPFC) components in our proposed model and would suggest these parameters to be relevant when assessing, monitoring or optimizing the effectiveness of psychotherapy for a given suicidality subtype. For example, we could think of the controllability parameter as reflecting the level of felt control over one's inner and outer life whereas the belief decay parameter could capture one's ability to unlearn maladaptive beliefs through new experiences, behavior or cognitive reappraisal (Zilverstand et al., 2017).

When it comes to pharmacotherapy, sub-anesthetic doses of ketamine, a N-methyl-D-aspartate receptor (NMDAR) antagonist, is currently one of the most promising interventions for rapid reduction of STB, but only 55–60% of individuals respond with a complete remission (Wilkinson et al., 2018). The exact mechanism through which ketamine achieves its anti-suicidal and anti-depressant effects is still not fully understood (Riggs and Gould, 2021). Many hypotheses emphasize the importance of increased α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPAR) signalling, its involvement in bottom-up information transmission and a consequent increase in synaptic and spine plasticity (Zanos and Gould, 2018; Lengvenyte et al., 2019). Other recent in vivo microdialysis findings suggest ketamine-induced AMPAR signaling in LC and DRN as well as a subsequent release of NE and 5-HT in the mPFC to be necessary for the rapid antidepressant effects (López-Gil et al., 2019; Llamosas et al., 2019; Pham et al., 2017), also implicating prelimbic cortex (a homolog to Brodmann's area 32 in the vmPFC) (PL)-DRN involvement in stressor controllability (Amat et al., 2016; Dolzani et al., 2018). A recent review also highlights the ACC to be playing a

DOI: 10.5334/cpsy.80

key role in mediating ketamine's antidepressant effects (Alexander et al., 2021). The model we introduced here could help provide a more mechanistic understanding of how the changes in belief updating and possibly activity in these brain regions relate to reduced suicide risk.

Personalization of early interventions could also be improved by a more mechanistic understanding of sex differences as it relates to STB (Williams and Trainor, 2018). Females show a higher incidence of suicidal intent and suicide attempts, although the rate of completed suicides is much higher in males (2 to 5 times) (Freeman et al., 2017). Suicide risk factors have also been found to differ between the sexes (Oquendo et al., 2007). While multiple psychosocial factors are likely to be contributing to these differences (Canetto and Sakinofsky, 1998), sexual dimorphisms in the brain might play an important role as well (Pallayova et al., 2019). For example, structural and functional dimorphisms in the LC-NE system and its regulation by estrogen in females is associated with an increased susceptibility to hyperarousal (Bangasser et al., 2016), which itself has been linked to a higher risk of suicidal ideation (Steyn et al., 2013; Morabito et al., 2020; Dolsen et al., 2017). Preclinical studies also suggest important sex differences in how stressor controllability modulates stress reactivity. Unlike males, females do not seem to benefit from increased controllability, with the lack of engagement and structural plasticity within the PL-DRN pathway being a likely mechanism for these differences (Fallon et al., 2020). The model proposed here might help better understand how these differences impact stress reactivity and controllability, and how this affects response to ketamine as well as to other interventions (Fallon et al., 2020).

4.2 LIMITATIONS

While we have considered some of the most crucial neurocircuits and neuromodulatory systems at the overlap of stress response, aversive learning, behavioral control and STB, there remain other relevant regions to be considered (Schmaal et al., 2020; Lengvenyte et al., 2019). Of particular importance may be the lateral habenula (LHb), an epithalamic nucleus acting as a relay hub between forebrain and midbrain structures and playing a significant role in learning from non-rewarding and aversive experiences (Matsumoto and Hikosaka, 2009). The LHb is involved in stressor controllability effects via the DRN-5-HT system (Metzger et al., 2017) and is one of the locations targeted by ketamine that mediates the anti-depressant effects (Zanos and Gould, 2018; Yang et al., 2018a; Shepard et al., 2018). LHb activity has been associated with depressive symptoms of helplessness, anhedonia, and excessive negative focus (Yang et al., 2018b), while a recent study also reported higher resting state functional connectivity between LHb and several brain regions, including the amygdala, to be associated with STB independently of depressive symptoms (Ambrosi et al., 2019).

It is worth emphasizing that while a close consideration of the networks implicated in STB informed the construction of the model proposed here, the implementation of the model is not at the level of neural dynamics but rather at the level of higher-order computational mechanisms underwritten by such dynamics (cf. Marr's levels of analysis (Marr and Poggio, 1976)). This means that the model variables might not necessarily neatly map onto distinct elements of the neurocircuitry but might interact with several other factors. For example, while we regard parameter c in the prior over outcomes to correspond to stress sensitivity and Amy activation, we could imagine other factors contributing to dispreference of the aversive outcome beyond its aversiveness per se, such as contextual factors relating to task engagement and a general motivation to do well in the task. Similarly, controllability threshold, w_0, might reflect a combined influence of changes in vmPFC activation, its connectivity to the DRN, connectivity from the DRN to the Amy or even the LHb and the effects it exerts on the DRN-5-HT system. Future work, including modelling of the neural dynamics and gathering empirical data, will help clarify these relationships.

Related to the above limitations, it is also important to reiterate that the hypothesis space presented in this paper serves as a proof of concept and is not intended to be exhaustive. The emergence of STB risk factors in different contexts is most likely to involve other variables that we have not yet considered. Furthermore, our simulations explored only the simplest scenarios of varying one parameter at a time. Considering how these parameters interact provides another layer of complexity. For example, we could expect different subtypes of STB to be related not to a single parameter, but to a unique combination of multiple parameters, forming distinct clusters

within the multidimensional parameter space. Future work with empirical data will allow for the further refinement of the hypothesis space and the delineation of different STB subtypes.

Furthermore, in the work presented here we have not explicitly addressed the distinction between suicide ideators and suicide attempters. Recent accounts of suicidality argue that suicidal ideation and the progression from ideation to attempts should be treated as separate processes (Van Orden et al., 2010; Klonsky and May, 2015; O'Connor and Kirtley, 2018; Bryan et al., 2020; Klonsky et al., 2018). The active inference framework might be well suited to study these distinctions as it explicitly models and factorizes inferences about the states of the world (cf. suicidal ideation) and action selection (cf. suicide attempt). This will be more thoroughly explored in future work.

Our proposed model and the insights it provides is also limited by the behavioral task to which the model was applied. In the task considered here, the stimulus is completely unambiguous and there is only one decision per trial to make. Notably, in the special case when outcomes unambiguously specify hidden states, active inference reduces to a simpler KL-control model (Friston et al., 2015), and is similar to model-based reinforcement learning models that are driven by reward maximization (see **Eq. (12)** in **S1 Appendix: full mathematical details of the model** for more detailed explanation). Introducing sensory uncertainty and multiple decisions would allow for the utilization of the unique aspects of active inference, namely epistemic action – a goal-directed sampling of information Friston et al. (2015). This would provide a more ecologically valid context to study the relationship between information sampling dynamics and STB. Such tasks would allow us to capture other phenomena relevant for STB, such as aversive generalization (how specific aversive events lead to negative beliefs about the world), its relationship to trauma, its effects on reduced problem-solving abilities (i.e. planning) and its influence on biases towards escape strategies (Linson and Friston, 2019; Linson et al., 2020).

Finally, it also important to point out that most of the model parameters that we focused on in this paper are not unique to the active inference framework. To explain the active-escape bias phenomenon, we had to introduce additional parameters and computations. Namely, we introduced the dependency of the learning rate on outcome values and the dependency of outcome values on controllability, with controllability being another addition in itself. Such a modelling approach where additional parameters and computations are added on top of the existing ones is not uncommon and has been applied in many reinforcement learning modelling approaches focused on Pavlovian and instrumental control mechanisms (e.g., Dorfman and Gershman, 2019; Guitart-Masip et al., 2012; Mkrtchian et al., 2017; Na et al., 2021; Grossman et al., 2021). However, it contrasts with most of the developments in active inference, where model extensions are derived from the free energy functional. While we justified the introduction of these computations by relying on a large body of literature investigating the mechanisms of interest, a more principled derivation of computations capturing these mechanisms might be possible.

DATA ACCESSIBILITY STATEMENTS

General Matlab code implementing Active Inference can be found at *https://www.fil.ion.ucl.ac.uk/spm/software/spm12/*. Code from this toolbox (*spm_mdp_VB.m*) was modified to perform the simulations presented in this paper and can be found at: *https://github.com/frank-pk/STB_AEGNG_AI*.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **S1 Appendix.** full mathematical details of the model. DOI: *https://doi.org/10.5334/cpsy.80.s1*

ACKNOWLEDGEMENTS

We acknowledge the generous support from the Krembil Foundation. We are also thankful to Dr. Venkat Bhat for clinical feedback and to Colleen Charlton for stylistic suggestions.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Povilas Karvelis: literature review, model implementation, visualization, writing of the first draft, writing and editing. Andreea O. Diaconescu: conceptualization, methodology, supervision, writing and editing.

AUTHOR AFFILIATIONS

Povilas Karvelis *orcid.org/0000-0001-7469-5624*
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada

Andreea O. Diaconescu *orcid.org/0000-0002-3633-9757*
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, Ontario, Canada; University of Toronto, Department of Psychiatry, Toronto, Ontario, Canada; Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada; Department of Psychology, University of Toronto, Toronto, ON, Canada

REFERENCES

Akam, T., Rodrigues-Vaz, I., Marcelo, I., Zhang, X., Pereira, M., Oliveira, R. F., Dayan, P., & **Costa, R. M.** (2021). The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron*, *109*(1), 149–163. DOI: *https://doi.org/10.1016/j.neuron.2020.10.013*

Alarcón, G., Sauder, M., Teoh, J. Y., Forbes, E. E., & **Quevedo, K.** (2019). Amygdala functional connectivity during self-face processing in depressed adolescents with recent suicide attempt. *Journal of the American Academy of Child & Adolescent Psychiatry*, *58*(2), 221–231. DOI: *https://doi.org/10.1016/j.jaac.2018.06.036*

Alexander, L., Jelen, L. A., Mehta, M. A., & **Young, A. H.** (2021). The anterior cingulate cortex as a key locus of ketamine's antidepressant action. *Neuroscience Biobehavioral Reviews*. DOI: *https://doi.org/10.1016/j.neubiorev.2021.05.003*

Amat, J., Dolzani, S. D., Tilden, S., Christianson, J. P., Kubala, K. H., Bartholomay, K., Sperr, K., Ciancio, N., Watkins, L. R., & **Maier, S. F.** (2016). Previous ketamine produces an enduring blockade of neurochemical and behavioral effects of uncontrollable stress. *Journal of Neuroscience, 36*(1), 153–161. DOI: *https://doi.org/10.1523/JNEUROSCI.3114-15.2016*

Ambrosi, E., Arciniegas, D. B., Curtis, K. N., Patriquin, M. A., Spalletta, G., Sani, G., Frueh, B. C., Fowler, J. C., Madan, A., & **Salas, R.** (2019). Resting-state functional connectivity of the habenula in mood disorder patients with and without suicide-related behaviors. *The Journal of neuropsychiatry and clinical neurosciences, 31*(1), 49–56. DOI: *https://doi.org/10.1176/appi.neuropsych.17120351*

Arango, V., Underwood, M. D., Boldrini, M., Tamir, H., Kassir, S. A., Hsiung, S.-c., Chen, J. J., & **Mann, J. J.** (2001). Serotonin 1a receptors, serotonin transporter binding and serotonin transporter mrna expression in the brainstem of depressed suicide victims. *Neuropsychopharmacology*, *25*(6), 892–903. DOI: *https://doi.org/10.1016/S0893-133X(01)00310-4*

Bach, H., Huang, Y.-Y., Underwood, M. D., Dwork, A. J., Mann, J. J., & **Arango, V.** (2014). Elevated serotonin and 5-hiaa in the brainstem and lower serotonin turnover in the prefrontal cortex of suicides. *Synapse, 68*(3), 127–130. DOI: *https://doi.org/10.1002/syn.21695*

Baek, K., Kwon, J., Chae, J.-H., Chung, Y. A., Kralik, J. D., Min, J.-A., Huh, H., Choi, K. M., Jang, K.-I., Lee, N.-B., et al. (2017). Heightened aversion to risk and loss in depressed patients with a suicide attempt history. *Scientific reports*, *7*(1), 1–13. DOI: *https://doi.org/10.1038/s41598-017-10541-5*

Balcioglu, Y. H., & **Kose, S.** (2018). Neural substrates of suicide and suicidal behaviour: from a neuroimaging perspective. *Psychiatry and Clinical Psychopharmacology, 28*(3), 314–328. DOI: *https://doi.org/10.1080/24750573.2017.1420378*

Bangasser, D. A., Wiersielis, K. R., & **Khantsis, S.** (2016). Sex differences in the locus coeruleusnorepinephrine system and its regulation by stress. *Brain research, 1641*, 177–188. DOI: *https://doi.org/10.1016/j.brainres.2015.11.021*

Baumeister, R. F. (1990). Suicide as escape from self. *Psychological review, 97*(1), 90. DOI: *https://doi.org/10.1037/0033-295X.97.1.90*

Bernanke, J., Stanley, B., & **Oquendo, M.** (2017). Toward fine-grained phenotyping of suicidal behavior: the role of suicidal subtypes. *Molecular psychiatry, 22*(8), 1080–1081. DOI: *https://doi.org/10.1038/mp.2017.123*

Bouret, S., & **Sara, S. J.** (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in neurosciences, 28*(11), 574–582. DOI: *https://doi.org/10.1016/j.tins.2005.09.002*

Bridge, J. A., Reynolds, B., McBee-Strayer, S. M., Sheftall, A. H., Ackerman, J., Stevens, J., Mendoza, K., Campo, J. V., & **Brent, D. A.** (2015). Impulsive aggression, delay discounting, and adolescent suicide attempts: effects of current psychotropic medication use and family history of suicidal behavior. *Journal of child and adolescent psychopharmacology, 25*(2), 114–123. DOI: *https://doi.org/10.1089/cap.2014.0042*

Briggs, S., Netuveli, G., Gould, N., Gkaravella, A., Gluckman, N. S., Kangogyere, P., Farr, R., Goldblatt, M. J., & **Lindner, R.** (2019). The effectiveness of psychoanalytic/psychodynamic psychotherapy for reducing suicide attempts and self-harm: systematic review and meta-analysis. *The British Journal of Psychiatry, 214*(6), 320–328. DOI: *https://doi.org/10.1192/bjp.2019.33*

Brown, V. M., Wilson, J., Hallquist, M. N., Szanto, K., & **Dombrovski, A. Y.** (2020). Ventromedial prefrontal value signals and functional connectivity during decision-making in suicidal behavior and impulsivity. *Neuropsychopharmacology, 45*(6), 1034–1041. DOI: *https://doi.org/10.1038/s41386-020-0632-0*

Bryan, C. J., Butner, J. E., May, A. M., Rugo, K. F., Harris, J. A., Oakey, D. N., Rozek, D. C., & **Bryan, A. O.** (2020). Nonlinear change processes and the emergence of suicidal behavior: A conceptual model based on the fluid vulnerability theory of suicide. *New ideas in psychology, 57*, 100758. DOI: *https://doi.org/10.1016/j.newideapsych.2019.100758*

Canetto, S. S., & **Sakinofsky, I.** (1998). The gender paradox in suicide. *Suicide and Life-Threatening Behavior, 28*(1), 1–23.

Chandler, D. J., Jensen, P., McCall, J. G., Pickering, A. E., Schwarz, L. A., & **Totah, N. K.** (2019). Redefining noradrenergic neuromodulation of behavior: impacts of a modular locus coeruleus architecture. *Journal of Neuroscience, 39*(42), 8239–8249. DOI: *https://doi.org/10.1523/JNEUROSCI.1164-19.2019*

Chesney, E., Goodwin, G. M., & **Fazel, S.** (2014). Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World psychiatry, 13*(2), 153–160. DOI: *https://doi.org/10.1002/wps.20128*

Clark, J. E., Watson, S., & **Friston, K. J.** (2018). What is mood? a computational perspective. *Psychological Medicine, 48*(14), 2277–2284. DOI: *https://doi.org/10.1017/S0033291718000430*

Clewett, D., Schoeke, A., & **Mather, M.** (2014). Locus coeruleus neuromodulation of memories encoded during negative or unexpected action outcomes. *Neurobiology of Learning and Memory, 111*, 65–70. DOI: *https://doi.org/10.1016/j.nlm.2014.03.006*

Collins, K. A., Mendelsohn, A., Cain, C. K., & **Schiller, D.** (2014). Taking action in the face of threat: neural synchronization predicts adaptive coping. *Journal of Neuroscience, 34*(44), 14733–14738. DOI: *https://doi.org/10.1523/JNEUROSCI.2152-14.2014*

Constant, A., Hesp, C., Davey, C. G., Friston, K. J., & **Badcock, P. B.** (2021). Why depressed mood is adaptive: A numerical proof of principle for an evolutionary systems theory of depression. *Computational psychiatry (Cambridge, Mass.), 5*(1), 60. DOI: *https://doi.org/10.5334/cpsy.70*

Cook, J. L., Swart, J. C., Froböse, M. I., Diaconescu, A. O., Geurts, D. E., Den Ouden, H. E., & **Cools, R.** (2019). Catecholaminergic modulation of meta-learning. *Elife, 8*, e51439. DOI: *https://doi.org/10.7554/eLife.51439.sa2*

Csifcsák, G., Melsæter, E., & **Mittner, M.** (2020). Intermittent absence of control during reinforcement learning interferes with pavlovian bias in action selection. *Journal of Cognitive Neuroscience, 32*(4), 646–663. DOI: *https://doi.org/10.1162/jocn_a_01515*

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & **Friston, K.** (2020). Active inference on discrete state-spaces: a synthesis. *Journal of Mathematical Psychology, 99*, 102447. DOI: *https://doi.org/10.1016/j.jmp.2020.102447*

Ding, Y., Lawrence, N., Olié, E., Cyprien, F., Le Bars, E., Bonafe, A., Phillips, M., Courtet, P., & **Jollant, F.** (2015). Prefrontal cortex markers of suicidal vulnerability in mood disorders: a model-based structural neuroimaging study with a translational perspective. *Translational psychiatry, 5*(2), e516–e516. DOI: *https://doi.org/10.1038/tp.2015.1*

Dolsen, M. R., Cheng, P., Arnedt, J. T., Swanson, L., Casement, M. D., Kim, H. S., Goldschmied, J. R., Hoffmann, R. F., Armitage, R., & **Deldin, P. J.** (2017). Neurophysiological correlates of suicidal ideation in major depressive disorder: hyperarousal during sleep. *Journal of affective disorders, 212*, 160–166. DOI: *https://doi.org/10.1016/j.jad.2017.01.025*

Dolzani, S., Baratta, M., Moss, J., Leslie, N., Tilden, S., Sørensen, A., Watkins, L., Lin, Y., & **Maier, S.** (2018). Inhibition of a descending prefrontal circuit prevents ketamine-induced stress resilience in females. *Eneuro, 5*(1). DOI: *https://doi.org/10.1523/ENEURO.0025-18.2018*

Dombrovski, A. Y., Clark, L., Siegle, G. J., Butters, M. A., Ichikawa, N., Sahakian, B. J., & **Szanto, K.** (2010). Reward/punishment reversal learning in older suicide attempters. *American Journal of Psychiatry, 167*(6), 699–707. DOI: *https://doi.org/10.1176/appi.ajp.2009.09030407*

Dombrovski, A. Y., & **Hallquist, M. N.** (2017). The decision neuroscience perspective on suicidal behavior: evidence and hypotheses. *Current opinion in psychiatry, 30*(1), 7. DOI: *https://doi.org/10.1097/YCO.0000000000000297*

Dombrovski, A. Y., & **Hallquist, M. N.** (2021). Search for solutions, learning, simulation, and choice processes in suicidal behavior. *Wiley Interdisciplinary Reviews: Cognitive Science*, page e1561. DOI: *https://doi.org/10.31234/osf.io/ezjt9*

Dombrovski, A. Y., Hallquist, M. N., Brown, V. M., Wilson, J., & **Szanto, K.** (2019). Value-based choice, contingency learning, and suicidal behavior in mid-and late-life depression. *Biological psychiatry, 85*(6), 506–516. DOI: *https://doi.org/10.1016/j.biopsych.2018.10.006*

Dorfman, H. M., & **Gershman, S. J.** (2019). Controllability governs the balance between pavlovian and instrumental action selection. *Nature communications, 10*(1), 1–8. DOI: *https://doi.org/10.1038/s41467-019-13737-7*

Eshel, N., & **Roiser, J. P.** (2010). Reward and punishment processing in depression. *Biological psychiatry, 68*(2), 118–124. DOI: *https://doi.org/10.1016/j.biopsych.2010.01.027*

Fallon, I. P., Tanner, M. K., Greenwood, B. N., & **Baratta, M. V.** (2020). Sex differences in resilience: Experiential factors and their mechanisms. *European Journal of Neuroscience, 52*(1), 2530–2547. DOI: *https://doi.org/10.1111/ejn.14639*

FitzGerald, T. H., Dolan, R. J., & **Friston, K. J.** (2014). Model averaging, optimal inference, and habit formation. *Frontiers in human neuroscience, 8*, 457. DOI: *https://doi.org/10.3389/fnhum.2014.00457*

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & **Nock, M. K.** (2017). Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin, 143*(2), 187. DOI: *https://doi.org/10.1037/bul0000084*

Freeman, A., Mergl, R., Kohls, E., Székely, A., Gusmao, R., Arensman, E., Koburger, N., Hegerl, U., & **Rummel-Kluge, C.** (2017). A cross-national study on gender differences in suicide intent. *BMC psychiatry, 17*(1), 1–11. DOI: *https://doi.org/10.1186/s12888-017-1398-8*

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews, 68*, 862–879. DOI: *https://doi.org/10.1016/j.neubiorev.2016.06.022*

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

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & **Dolan, R. J.** (2013). The anatomy of choice: active inference and agency. *Frontiers in human neuroscience, 7*, 598. DOI: *https://doi.org/10.3389/fnhum.2013.00598*

Gläscher, J., Daw, N., Dayan, P., & **O'Doherty, J. P.** (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*(4), 585–595. DOI: *https://doi.org/10.1016/j.neuron.2010.04.016*

Grahn, R. E., Hammack, S., Will, M., O'Connor, K., Deak, T., Sparks, P., Watkins, L., & **Maier, S.** (2002). Blockade of alpha1 adrenoreceptors in the dorsal raphe nucleus prevents enhanced conditioned fear and impaired escape performance following uncontrollable stressor exposure in rats. *Behavioural brain research, 134*(1–2), 387–392. DOI: *https://doi.org/10.1016/S0166-4328(02)00061-X*

Grossman, C. D., Bari, B. A., & **Cohen, J. Y.** (2021). Serotonin neurons modulate learning rate through uncertainty. *Current Biology*. DOI: *https://doi.org/10.1101/2020.10.24.353508*

Grueschow, M., Kleim, B., & **Ruff, C. C.** (2020). Role of the locus coeruleus arousal system in cognitive control. *Journal of Neuroendocrinology*, page e12890. DOI: *https://doi.org/10.1111/jne.12890*

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & **Dolan, R. J.** (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage, 62*(1), 154–166. DOI: *https://doi.org/10.1016/j.neuroimage.2012.04.024*

Harfmann, E. J., Rhyner, K. T., & **Ingram, R. E.** (2019). Cognitive inhibition and attentional biases in the affective go/no-go performance of depressed, suicidal populations. *Journal of affective disorders, 256*, 228–233. DOI: *https://doi.org/10.1016/j.jad.2019.05.022*

Hebb, D. O. (1955). Drives and the cns (conceptual nervous system). *Psychological review, 62*(4), 243. DOI: *https://doi.org/10.1037/h0041823*

Hiser, J., & **Koenigs, M.** (2018). The multifaceted role of the ventromedial prefrontal cortex in emotion, decision making, social cognition, and psychopathology. *Biological psychiatry, 83*(8), 638–647. DOI: *https://doi.org/10.1016/j.biopsych.2017.10.030*

Holmes, N. M., Marchand, A. R., & **Coutureau, E.** (2010). Pavlovian to instrumental transfer: a neurobehavioural perspective. *Neuroscience & Biobehavioral Reviews*, *34*(8), 1277–1295. DOI: *https://doi.org/10.1016/j.neubiorev.2010.03.007*

Holroyd, C. B., & **Yeung, N.** (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in cognitive sciences*, *16*(2), 122–128. DOI: *https://doi.org/10.1016/j.tics.2011.12.008*

Hrdina, P. D., Demeter, E., Vu, T. B., Sótónyi, P., & **Palkovits, M.** (1993). 5-ht uptake sites and 5-ht2 receptors in brain of antidepressant-free suicide victims/depressives: increase in 5-ht2 sites in cortex and amygdala. *Brain research, 614*(1–2), 37–44. DOI: *https://doi.org/10.1016/0006-8993(93)91015-K*

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

Isometsä, E. (2014). Suicidal behaviour in mood disorders-who, when, and why? *The Canadian Journal of Psychiatry, 59*(3), 120–130. DOI: *https://doi.org/10.1177/070674371405900303*

Iyengar, U., Snowden, N., Asarnow, J. R., Moran, P., Tranah, T., & **Ougrin, D.** (2018). A further look at therapeutic interventions for suicide attempts and self-harm in adolescents: an updated systematic review of randomized controlled trials. *Frontiers in psychiatry, 9*, 583. DOI: *https://doi.org/10.3389/fpsyt.2018.00583*

Jacobs, H. I., Priovoulos, N., Poser, B. A., Pagen, L. H., Ivanov, D., Verhey, F. R., & **Uludağ, K.** (2020). Dynamic behavior of the locus coeruleus during arousal-related memory processing in a multi-modal 7t fmri paradigm. *Elife, 9*, e52059. DOI: *https://doi.org/10.7554/eLife.52059.sa2*

Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., & **Nieuwenhuis, S.** (2016). Catecholaminergic regulation of learning rate in a dynamic environment. *PLoS Computational Biology, 12*(10), e1005171. DOI: *https://doi.org/10.1371/journal.pcbi.1005171*

Joffily, M., & **Coricelli, G.** (2013). Emotional valence and the free-energy principle. *PLoS Comput Biol, 9*(6), e1003094. DOI: *https://doi.org/10.1371/journal.pcbi.1003094*

Jolly, T., Trivedi, C., Adnan, M., Mansuri, Z., & **Agarwal, V.** (2021). Gambling in patients with major depressive disorder is associated with an elevated risk of suicide: Insights from 12-years of nationwide inpatient sample data. *Addictive behaviors, 118*, 106872. DOI: *https://doi.org/10.1016/j.addbeh.2021.106872*

Jylhä, P., Rosenström, T., Mantere, O., Suominen, K., Melartin, T., Vuorilehto, M., Holma, M., Riihimäki, K., Oquendo, M. A., Keltikangas-Järvinen, L., et al. (2016). Personality disorders and suicide attempts in unipolar and bipolar mood disorders. *Journal of affective disorders, 190*, 632–639. DOI: *https://doi.org/10.1016/j.jad.2015.11.006*

Kang, S.-G., Na, K.-S., Choi, J.-W., Kim, J.-H., Son, Y.-D., & **Lee, Y. J.** (2017). Resting-state functional connectivity of the amygdala in suicide attempters with major depressive disorder. *Progress in neuro-psychopharmacology and biological psychiatry, 77*, 222–227. DOI: *https://doi.org/10.1016/j.pnpbp.2017.04.029*

Karlsson, A., & **Håkansson, A.** (2018). Gambling disorder, increased mortality, suicidality, and associated comorbidity: A longitudinal nationwide register study. *Journal of behavioral addictions, 7*(4), 1091–1099. DOI: *https://doi.org/10.1556/2006.7.2018.112*

Kerr, D. L., McLaren, D. G., Mathy, R. M., & **Nitschke, J. B.** (2012). Controllability modulates the anticipatory response in the human ventromedial prefrontal cortex. *Frontiers in Psychology, 3*, 557. DOI: *https://doi.org/10.3389/fpsyg.2012.00557*

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & **Zubizarreta, J. R.** (2020). Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular psychiatry*, *25*(1), 168–179. DOI: *https://doi.org/10.1038/s41380-019-0531-0*

Klonsky, E. D., & **May, A. M.** (2015). The three-step theory (3st): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy, 8*(2), 114–129. DOI: *https://doi.org/10.1521/ijct.2015.8.2.114*

Klonsky, E. D., Saffer, B. Y., & **Bryan, C. J.** (2018). Ideation-to-action theories of suicide: a conceptual and empirical update. *Current opinion in psychology, 22*, 38–43. DOI: *https://doi.org/10.1016/j.copsyc.2017.07.020*

Köhler, S., Bär, K.-J., & **Wagner, G.** (2016). Differential involvement of brainstem noradrenergic and midbrain dopaminergic nuclei in cognitive control. *Human Brain Mapping, 37*(6), 2305–2318. DOI: *https://doi.org/10.1002/hbm.23173*

Lalovic, A., Wang, S., Keilp, J. G., Bowie, C. R., Kennedy, S. H., & **Rizvi, S. J.** (2022). A qualitative systematic review of neurocognition in suicide ideators and attempters: Implications for cognitive-based psychotherapeutic interventions. *Neuroscience & Biobehavioral Reviews, 132*, 92–109. DOI: *https://doi.org/10.1016/j.neubiorev.2021.11.007*

Large, M., Kaneson, M., Myles, N., Myles, H., Gunaratne, P., & **Ryan, C.** (2016). Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. *PloS one, 11*(6), e0156322. DOI: *https://doi.org/10.1371/journal.pone.0156322*

Lawson, R. P., Bisby, J., Nord, C. L., Burgess, N., & **Rees, G.** (2020). The computational, pharmacological, and physiological determinants of sensory learning under uncertainty. *Current Biology*.

Lengvenyte, A., Olié, E., & **Courtet, P.** (2019). Suicide has many faces, so does ketamine: a narrative review on ketamine's antisuicidal actions. *Current psychiatry reports, 21*(12), 1–10. DOI: *https://doi.org/10.1007/s11920-019-1108-y*

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

Linson, A., Parr, T., & **Friston, K. J.** (2020). Active inference, stressors, and psychological trauma: A neuroethological model of (mal) adaptive explore-exploit dynamics in ecological context. *Behavioural brain research, 380*, 112421. DOI: *https://doi.org/10.1016/j.bbr.2019.112421*

Liu, R. T., Cheek, S. M., & **Nestor, B. A.** (2016). Non-suicidal self-injury and life stress: A systematic meta-analysis and theoretical elaboration. *Clinical psychology review, 47*, 1–14. DOI: *https://doi.org/10.1016/j.cpr.2016.05.005*

Liu, R. T., Kleiman, E. M., Nestor, B. A., & **Cheek, S. M.** (2015). The hopelessness theory of depression: A quarter-century in review. *Clinical Psychology: Science and Practice, 22*(4), 345. DOI: *https://doi.org/10.1037/h0101732*

Llamosas, N., Perez-Caballero, L., Berrocoso, E., Bruzos-Cidon, C., Ugedo, L., & **Torrecilla, M.** (2019). Ketamine promotes rapid and transient activation of ampa receptor-mediated synaptic transmission in the dorsal raphe nucleus. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *88*, 243–252. DOI: *https://doi.org/10.1016/j.pnpbp.2018.07.022*

López-Gil, X., Jiménez-Sánnchez, L., Campa, L., Castro, E., Frago, C., & **Adell, A.** (2019). Role of serotonin and noradrenaline in the rapid antidepressant action of ketamine. *ACS chemical neuroscience, 10*(7), 3318–3326. DOI: *https://doi.org/10.1021/acschemneuro.9b00288*

Maier, S. F., & **Seligman, M. E.** (2016). Learned helplessness at fifty: Insights from neuroscience. *Psychological review, 123*(4), 349. DOI: *https://doi.org/10.1037/rev0000033*

Mann, J. J., Currier, D., Stanley, B., Oquendo, M. A., Amsel, L. V., & **Ellis, S. P.** (2006). Can biological tests assist prediction of suicide in mood disorders? *International Journal of Neuropsychopharmacology, 9*(4), 465–474. DOI: *https://doi.org/10.1017/S1461145705005687*

Mann, J. J., Huang, Y.-y., Underwood, M. D., Kassir, S. A., Oppenheim, S., Kelly, T. M., Dwork, A. J., & **Arango, V.** (2000). A serotonin transporter gene promoter polymorphism (5-httlpr) and prefrontal cortical binding in major depression and suicide. *Archives of general psychiatry, 57*(8), 729–738. DOI: *https://doi.org/10.1001/archpsyc.57.8.729*

Mann, J. J., & **Rizk, M. M.** (2020). A brain-centric model of suicidal behavior. *American journal of psychiatry*, *177*(10), 902–916. DOI: *https://doi.org/10.1176/appi.ajp.2020.20081224*

Marr, D., & **Poggio, T.** (1976). From understanding computation to understanding neural circuitry.

Mathews, A., & **MacLeod, C.** (2005). Cognitive vulnerability to emotional disorders. *Annu. Rev. Clin. Psychol., 1*, 167–195. DOI: *https://doi.org/10.1146/annurev.clinpsy.1.102803.143916*

Matsumoto, M., & **Hikosaka, O.** (2009). Representation of negative motivational value in the primate lateral habenula. *Nature neuroscience, 12*(1), 77. DOI: *https://doi.org/10.1038/nn.2233*

Matsumoto, M., Matsumoto, K., Abe, H., & **Tanaka, K.** (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature neuroscience*, *10*(5), 647–656. DOI: *https://doi.org/10.1038/nn1890*

May, A. M., Pachkowski, M. C., & **Klonsky, E. D.** (2020). Motivations for suicide: Converging evidence from clinical and community samples. *Journal of psychiatric research, 123*, 171–177. DOI: *https://doi.org/10.1016/j.jpsychires.2020.02.010*

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological inquiry*, *1*(2), 108–141. DOI: *https://doi.org/10.1207/s15327965pli0102_1*

Mérelle, S., Foppen, E., Gilissen, R., Mokkenstorm, J., Cluitmans, R., & **Van Ballegooijen, W.** (2018). Characteristics associated with non-disclosure of suicidal ideation in adults. *International journal of environmental research and public health, 15*(5), 943. DOI: *https://doi.org/10.3390/ijerph15050943*

Messina, I., Sambin, M., Beschoner, P., & **Viviani, R.** (2016). Changing views of emotion regulation and neurobiological models of the mechanism of action of psychotherapy. *Cognitive, Affective, & Behavioral Neuroscience, 16*(4), 571–587. DOI: *https://doi.org/10.3758/s13415-016-0440-5*

Metzger, M., Bueno, D., & **Lima, L. B.** (2017). The lateral habenula and the serotonergic system. *Pharmacology Biochemistry and Behavior, 162*, 22–28. DOI: *https://doi.org/10.1016/j.pbb.2017.05.007*

Miller, A. B., McLaughlin, K. A., Busso, D. S., Brueck, S., Peverill, M., & **Sheridan, M. A.** (2018). Neural correlates of emotion regulation and adolescent suicidal ideation. *Biological psychiatry: cognitive neuroscience and neuroimaging, 3*(2), 125–132. DOI: *https://doi.org/10.1016/j.bpsc.2017.08.008*

Millner, A. J., den Ouden, H. E., Gershman, S. J., Glenn, C. R., Kearns, J. C., Bornstein, A. M., Marx, B. P., Keane, T. M., & **Nock, M. K.** (2019). Suicidal thoughts and behaviors are associated with an increased decision-making bias for active responses to escape aversive states. *Journal of abnormal psychology, 128*(2), 106. DOI: *https://doi.org/10.1037/abn0000395*

Millner, A. J., Gershman, S. J., Nock, M. K., & **den Ouden, H. E.** (2018). Pavlovian control of escape and avoidance. *Journal of Cognitive Neuroscience, 30*(10), 1379–1390. DOI: *https://doi.org/10.1162/jocn_a_01224*

Millner, A. J., Robinaugh, D. J., & **Nock, M. K.** (2020). Advancing the understanding of suicide: the need for formal theory and rigorous descriptive research. *Trends in cognitive sciences*. DOI: *https://doi.org/10.1016/j.tics.2020.06.007*

Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., & **Robinson, O. J.** (2017). Modeling avoidance in mood and anxiety disorders using reinforcement learning. *Biological psychiatry, 82*(7), 532–539. DOI: *https://doi.org/10.1016/j.biopsych.2017.01.017*

Monkul, E., Hatch, J. P., Nicoletti, M. A., Spence, S., Brambilla, P., Lacerda, A. L. T. d., Sassi, R. B., Mallinger, A., Keshavan, M., & **Soares, J. C.** (2007). Fronto-limbic brain structures in suicidal and non-suicidal female patients with major depressive disorder. *Molecular psychiatry, 12*(4), 360–366. DOI: *https://doi.org/10.1038/sj.mp.4001919*

Morabito, D. M., Boffa, J. W., Bedford, C. E., Chen, J. P., & **Schmidt, N. B.** (2020). Hyperarousal symptoms and perceived burdensomeness interact to predict suicidal ideation among trauma-exposed individuals. *Journal of psychiatric research, 130*, 218–223. DOI: *https://doi.org/10.1016/j.jpsychires.2020.07.029*

Na, S., Chung, D., Hula, A., Perl, O., Jung, J., Hein, M., Blackmore, S., Fiore, V. G., Dayan, P., & **Gu, X.** (2021). Humans use forward thinking to exploit social controllability. *Elife, 10*, e64983. DOI: *https://doi.org/10.7554/eLife.64983.sa2*

Naghavi, M., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abera, S. F., Aboyans, V., Adetokunboh, O., Afshin, A., Agrawal, A., et al. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet, 390*(10100), 1151–1210. DOI: *https://doi.org/10.1016/S0140-6736(17)32152-9*

Nair, A., Rutledge, R. B., & **Mason, L.** (2020). Under the hood: using computational psychiatry to make psychological therapies more mechanism-focused. *Frontiers in psychiatry, 11*, 140. DOI: *https://doi.org/10.3389/fpsyt.2020.00140*

Nock, M. K., Hwang, I., Sampson, N. A., & **Kessler, R. C.** (2010). Mental disorders, comorbidity and suicidal behavior: results from the national comorbidity survey replication. *Molecular psychiatry, 15*(8), 868–876. DOI: *https://doi.org/10.1038/mp.2009.29*

O'Connor, R. C., & **Kirtley, O. J.** (2018). The integrated motivational-volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1754), 20170268. DOI: *https://doi.org/10.1098/rstb.2017.0268*

Olié, E., Ding, Y., Le Bars, E., de Champeur, N. M., Mura, T., Bonafé, A., Courtet, P., & **Jollant, F.** (2015). Processing of decision-making and social threat in patients with history of suicidal attempt: A neuroimaging replication study. *Psychiatry Research: Neuroimaging, 234*(3), 369–377. DOI: *https://doi.org/10.1016/j.pscychresns.2015.09.020*

Oquendo, M. A., Bongiovi-Garcia, M. E., Galfalvy, H., Goldberg, P. H., Grunebaum, M. F., Burke, A. K., & **Mann, J. J.** (2007). Sex differences in clinical predictors of suicidal acts after major depression: a prospective study. *American Journal of Psychiatry*, *164*(1), 134–141. DOI: *https://doi.org/10.1176/ajp.2007.164.1.134*

Oquendo, M. A., Galfalvy, H. C., Choo, T.-H., Kandlur, R., Burke, A. K., Sublette, M. E., Miller, J. M., Mann, J. J., & **Stanley, B. H.** (2020). Highly variable suicidal ideation: a phenotypic marker for stress induced suicide risk. *Molecular psychiatry*, pages 1–8. DOI: *https://doi.org/10.1038/s41380-020-0819-0*

Oquendo, M. A., Sullivan, G. M., Sudol, K., Baca-Garcia, E., Stanley, B. H., Sublette, M. E., & **Mann, J. J.** (2014). Toward a biosignature for suicide. *American Journal of Psychiatry, 171*(12), 1259–1277. DOI: *https://doi.org/10.1176/appi.ajp.2014.14020194*

Ougrin, D., Tranah, T., Stahl, D., Moran, P., & **Asarnow, J. R.** (2015). Therapeutic interventions for suicide attempts and self-harm in adolescents: systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(2), 97–107. DOI: *https://doi.org/10.1016/j.jaac.2014.10.009*

Pallayova, M., Brandeburova, A., & **Tokarova, D.** (2019). Update on sexual dimorphism in brain structure-function interrelationships: A literature review. *Applied psychophysiology and biofeedback*, 44(4), 271–284. DOI: *https://doi.org/10.1007/s10484-019-09443-1*

Pezzulo, G., Rigoli, F., & **Friston, K.** (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology, 134*, 17–35. DOI: *https://doi.org/10.1016/j.pneurobio.2015.09.001*

Pham, T., Mendez-David, I., Defaix, C., Guiard, B., Tritschler, L., David, D., & **Gardier, A.** (2017). Ketamine treatment involves medial prefrontal cortex serotonin to induce a rapid antidepressant-like activity in balb/cj mice. *Neuropharmacology*, *112*, 198–209. DOI: *https://doi.org/10.1016/j.neuropharm.2016.05.010*

Poe, G. R., Foote, S., Eschenko, O., Johansen, J. P., Bouret, S., Aston-Jones, G., Harley, C. W., Manahan-Vaughan, D., Weinshenker, D., Valentino, R., et al. (2020). Locus coeruleus: a new look at the blue spot. *Nature Reviews Neuroscience*, pages 1–16. DOI: *https://doi.org/10.1038/s41583-020-0360-9*

Pudovkina, O. L., Cremers, T. I., & **Westerink, B. H.** (2003). Regulation of the release of serotonin in the dorsal raphe nucleus by α1 and α2 adrenoceptors. *Synapse, 50*(1), 77–82. DOI: *https://doi.org/10.1002/syn.10245*

Pulcu, E., & **Browning, M.** (2017). Affective bias as a rational response to the statistics of rewards and punishments. *Elife, 6*, e27879. DOI: *https://doi.org/10.7554/eLife.27879.018*

Pulcu, E., & **Browning, M.** (2019). The misestimation of uncertainty in affective disorders. *Trends in Cognitive Sciences*, *23*(10), 865–875. DOI: *https://doi.org/10.1016/j.tics.2019.07.007*

Richard-Devantoy, S., Berlim, M., & **Jollant, F.** (2014). A meta-analysis of neuropsychological markers of vulnerability to suicidal behavior in mood disorders. *Psychological medicine, 44*(8), 1663–1673. DOI: *https://doi.org/10.1017/S0033291713002304*

Riggs, L. M., & **Gould, T. D.** (2021). Ketamine and the future of rapid-acting antidepressants. *Annual Review of Clinical Psychology*, 17. DOI: *https://doi.org/10.1146/annurev-clinpsy-072120-014126*

Rizk, M. M., Choo, T.-H., Galfalvy, H., Biggs, E., Brodsky, B. S., Oquendo, M. A., Mann, J. J., & **Stanley, B.** (2019). Variability in suicidal ideation is associated with affective instability in suicide attempters with borderline personality disorder. *Psychiatry*, *82*(2), 173–178. DOI: *https://doi.org/10.1080/00332747.2019.1600219*

Rushworth, M. F., & **Behrens, T. E.** (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature neuroscience*, *11*(4), 389–397. DOI: *https://doi.org/10.1038/nn2066*

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & **Moran, R. J.** (2019). Locus coeruleus tracking of prediction errors optimises cognitive exibility: An active inference model. *PLoS computational biology, 15*(1), e1006267. DOI: *https://doi.org/10.1371/journal.pcbi.1006267*

Sastre-Buades, A., Alacreu-Crespo, A., Courtet, P., Baca-Garcia, E., & **Barrigon, M. L.** (2021). Decisionmaking in suicidal behavior: a systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews, 131*, 642–662. DOI: *https://doi.org/10.1016/j.neubiorev.2021.10.005*

Schmaal, L., van Harmelen, A.-L., Chatzi, V., Lippard, E. T., Toenders, Y. J., Averill, L. A., Mazure, C. M., & **Blumberg, H. P.** (2020). Imaging suicidal thoughts and behaviors: a comprehensive review of 2 decades of neuroimaging studies. *Molecular psychiatry, 25*(2), 408–427. DOI: *https://doi.org/10.1038/s41380-019-0587-x*

Sharot, T., & **Garrett, N.** (2016). Forming beliefs: Why valence matters. *Trends in cognitive sciences, 20*(1), 25–33. DOI: *https://doi.org/10.1016/j.tics.2015.11.002*

Shenhav, A., Botvinick, M. M., & **Cohen, J. D.** (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), 217–240. DOI: *https://doi.org/10.1016/j.neuron.2013.07.007*

Shepard, R. D., Langlois, L. D., Browne, C. A., Berenji, A., Lucki, I., & **Nugent, F. S.** (2018). Ketamine reverses lateral habenula neuronal dysfunction and behavioral immobility in the forced swim test following maternal deprivation in late adolescent rats. *Frontiers in synaptic neuroscience*, *10*, 39. DOI: *https://doi.org/10.3389/fnsyn.2018.00039*

Silvetti, M., Vassena, E., Abrahamse, E., & **Verguts, T.** (2018). Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS computational biology, 14*(8), e1006370. DOI: *https://doi.org/10.1371/journal.pcbi.1006370*

Smith, R., Badcock, P., & **Friston, K. J.** (2021). Recent advances in the application of predictive coding and active inference models within clinical neuroscience. *Psychiatry and Clinical Neurosciences*, *75*(1), 3–13. DOI: *https://doi.org/10.1111/pcn.13138*

Spoletini, I., Piras, F., Fagioli, S., Rubino, I. A., Martinotti, G., Siracusano, A., Caltagirone, C., & **Spalletta, G.** (2011). Suicidal attempts and increased right amygdala volume in schizophrenia. *Schizophrenia research*, *125*(1), 30–40. DOI: *https://doi.org/10.1016/j.schres.2010.08.023*

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., Degueldre, C., Luxen, A., Collette, F., & **Maquet, P.** (2006). The locus ceruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience, 26*(28), 7416–7423. DOI: *https://doi.org/10.1523/JNEUROSCI.1001-06.2006*

Steyn, R., Vawda, N., Wyatt, G. E., Williams, J., & **Madu, S.** (2013). Posttraumatic stress disorder diagnostic criteria and suicidal ideation in a south african police sample. *African journal of psychiatry, 16*(1), 19–22. DOI: *https://doi.org/10.4314/ajpsy.v16i1.3*

Stolz, D. S., Müller-Pinzler, L., Krach, S., & **Paulus, F. M.** (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature communications, 11*(1), 1–13. DOI: *https://doi.org/10.1038/s41467-020-14800-4*

Szanto, K., Clark, L., Hallquist, M., Vanyukov, P., Crockett, M., & **Dombrovski, A. Y.** (2014). The cost of social punishment and high-lethality suicide attempts in the second half of life. *Psychology and aging, 29*(1), 84. DOI: *https://doi.org/10.1037/a0035339*

Tervo, D. G., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & **Karpova, A. Y.** (2014). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell, 159*(1), 21–32. DOI: *https://doi.org/10.1016/j.cell.2014.08.037*

Uematsu, A., Tan, B. Z., Ycu, E. A., Cuevas, J. S., Koivumaa, J., Junyent, F., Kremer, E. J., Witten, I. B., Deisseroth, K., & **Johansen, J. P.** (2017). Modular organization of the brainstem noradrenaline system coordinates opposing learning states. *Nature neuroscience, 20*(11), 1602. DOI: *https://doi.org/10.1038/nn.4642*

van Heeringen, K., & **Mann, J. J.** (2014). The neurobiology of suicide. *The Lancet Psychiatry, 1*(1), 63–72. DOI: *https://doi.org/10.1016/S2215-0366(14)70220-2*

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & **Joiner, T. E., Jr.** (2010). The interpersonal theory of suicide. *Psychological review*, *117*(2), 575. DOI: *https://doi.org/10.1037/a0018697*

Vanyukov, P. M., Szanto, K., Hallquist, M. N., Siegle, G. J., Reynolds, C. F., III, Forman, S. D., Aizenstein, H. J., & **Dombrovski, A. Y.** (2016). Paralimbic and lateral prefrontal encoding of reward value during intertemporal choice in attempted suicide. *Psychological medicine, 46*(2), 381. DOI: *https://doi.org/10.1017/S0033291715001890*

Vassena, E., Krebs, R. M., Silvetti, M., Fias, W., & **Verguts, T.** (2014). Dissociating contributions of acc and vmpfc in reward prediction, outcome, and choice. *Neuropsychologia, 59*, 112–123. DOI: *https://doi.org/10.1016/j.neuropsychologia.2014.04.019*

Verrocchio, M. C., Carrozzino, D., Marchetti, D., Andreasson, K., Fulcheri, M., & **Bech, P.** (2016). Mental pain and suicide: a systematic review of the literature. *Frontiers in Psychiatry, 7*, 108. DOI: *https://doi.org/10.3389/fpsyt.2016.00108*

Wagner, G., Koch, K., Schachtzabel, C., Schultz, C. C., Sauer, H., & **Schlösser, R. G.** (2011). Structural brain alterations in patients with major depressive disorder and high risk for suicide: evidence for a distinct neurobiological entity? *Neuroimage, 54*(2), 1607–1614. DOI: *https://doi.org/10.1016/j.neuroimage.2010.08.082*

Wang, K. S., & **Delgado, M. R.** (2019). Corticostriatal circuits encode the subjective value of perceived control. *Cerebral Cortex*, *29*(12), 5049–5060. DOI: *https://doi.org/10.1093/cercor/bhz045*

Wang, L., Zhao, Y., Edmiston, E. K., Womer, F. Y., Zhang, R., Zhao, P., Jiang, X., Wu, F., Kong, L., Zhou, Y., et al. (2020). Structural and functional abnormities of amygdala and prefrontal cortex in major depressive disorder with suicide attempts. *Frontiers in psychiatry, 10*, 923. DOI: *https://doi.org/10.3389/fpsyt.2019.00923*

Wanke, N., & **Schwabe, L.** (2020). Dissociable neural signatures of passive extinction and instrumental control over threatening events. *Social cognitive and affective neuroscience, 15*(6), 625–634. DOI: *https://doi.org/10.1093/scan/nsaa074*

Weinberg, I., Ronningstam, E., Goldblatt, M. J., Schechter, M., Wheelis, J., & **Maltsberger, J. T.** (2010). Strategies in treatment of suicidality: identification of common and treatment-specific interventions in empirically supported treatment manuals. *The Journal of clinical psychiatry*, *71*(6), 699–706. DOI: *https://doi.org/10.4088/JCP.08m04840blu*

Wilkinson, S. T., Ballard, E. D., Bloch, M. H., Mathew, S. J., Murrough, J. W., Feder, A., Sos, P., Wang, G., Zarate, C. A., Jr., & **Sanacora, G.** (2018). The effect of a single dose of intravenous ketamine on suicidal ideation: a systematic review and individual participant data meta-analysis. *American journal of psychiatry, 175*(2), 150–158. DOI: *https://doi.org/10.1176/appi.ajp.2017.17040472*

Willeumier, K., Taylor, D. V., & **Amen, D. G.** (2011). Decreased cerebral blood flow in the limbic and prefrontal cortex using spect imaging in a cohort of completed suicides. *Translational psychiatry, 1*(8), e28–e28. DOI: *https://doi.org/10.1038/tp.2011.28*

Williams, A. V., & **Trainor, B. C.** (2018). The impact of sex as a biological variable in the search for novel antidepressants. *Frontiers in neuroendocrinology*, *50*, 107–117. DOI: *https://doi.org/10.1016/j.yfrne.2018.05.003*

Yang, Y., Cui, Y., Sang, K., Dong, Y., Ni, Z., Ma, S., & **Hu, H.** (2018a). Ketamine blocks bursting in the lateral habenula to rapidly relieve depression. *Nature, 554*(7692), 317–322. DOI: *https://doi.org/10.1038/nature25509*

Yang, Y., Wang, H., Hu, J., & **Hu, H.** (2018b). Lateral habenula in the pathophysiology of depression. *Current opinion in neurobiology*, *48*, 90–96. DOI: *https://doi.org/10.1016/j.conb.2017.10.024*

Yerkes, R. M., Dodson, J. D., et al. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, pages 27–41.

Yu, A. J., & **Dayan, P.** (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681–692. DOI: *https://doi.org/10.1016/j.neuron.2005.04.026*

Zanos, P., & **Gould, T. D.** (2018). Mechanisms of ketamine action as an antidepressant. *Molecular psychiatry*, *23*(4), 801–811. DOI: *https://doi.org/10.1038/mp.2017.255*

Zilverstand, A., Parvaz, M. A., & **Goldstein, R. Z.** (2017). Neuroimaging cognitive reappraisal in clinical populations to define neural targets for enhancing emotion regulation. A systematic review. *Neuroimage*, *151*, 105–116. DOI: *https://doi.org/10.1016/j.neuroimage.2016.06.009*

TO CITE THIS ARTICLE:
Karvelis, P., & Diaconescu, A. O. (2022). A Computational Model of Hopelessness and Active-Escape Bias in Suicidality. *Computational Psychiatry*, 6(1), pp. 34–59. DOI: *https://doi.org/10.5334/cpsy.80*

Submitted: 07 September 2021
Accepted: 15 February 2022
Published: 31 March 2022

Computational Psychiatry is a peer-reviewed open access journal published by Ubiquity Press.