JNeurosci THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Under threat weaker evidence is required to reach undesirable conclusions

https://doi.org/10.1523/JNEUROSCI.3194-20.2021

Cite as: J. Neurosci 2021; 10.1523/JNEUROSCI.3194-20.2021

Received: 21 December 2020
Revised: 24 March 2021
Accepted: 5 May 2021

Under threat weaker evidence is required to reach undesirable conclusions

Abbreviated Title: Evidence Accumulation under Threat

Authors: Laura K. Globig[1,2] *, Kristin Witte[3,4], Gloria Feng[5] & Tali Sharot[1,2]*

[1] Affective Brain Lab, Department of Experimental Psychology, University College London, London, WC1H 0AP, UK

[2]The Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, WC1H 0AP, UK

[3] University of Tübingen, Tübingen, 72074, Germany

[4] The Max Planck Institute for Biological Cybernetics, Tübingen, 72076, Germany

[5] Yale University, New Haven, CT 06520-8205, USA

*Correspondence to: laura.globig.15@ucl.ac.uk, t.sharot@ucl.ac.uk

Number of pages: 28

Number of figures: 6

Number of tables: 4

Number of Words for Abstract: 206

Number of Words for Introduction: 650

Number of Words for Discussion: 728

Conflict of Interest Statement.
No competing interests declared.

Acknowledgments.
The research was funded by a Wellcome Trust Senior Research Fellowship 214268/Z/18/Z to TS. We thank Bastien Blain, Irene Cogliati Dezza, Moshe Glickman, Christopher Kelly, Gaia Molinaro, Christina Maher, Valentina Vellani, and Sarah Zheng for providing comments on a previous version of the manuscript.

Abstract. Critical decisions, such as in domains ranging from medicine to finance, are often made under threatening circumstances that elicit stress and anxiety. The negative effects of such reactions on learning and decision-making have been repeatedly underscored. In contrast, here we show that perceived threat alters the process by which evidence is accumulated in a way that may be adaptive. Participants (n = 91) completed a sequential evidence sampling task in which they were incentivized to accurately judge whether they were in a desirable state, which was associated with greater rewards than losses, or an undesirable state, which was associated with greater losses than rewards. Prior to the task participants in the 'threat group' experienced a social-threat manipulation. Results show that perceived threat led to a reduction in the strength of evidence required to reach an undesirable judgement. Computational modelling revealed this was due to an increase in the relative rate by which negative information was accumulated. The effect of the threat manipulation was global, as the alteration to evidence accumulation was observed for information which was not directly related to the cause of the threat. Requiring weaker evidence to reach undesirable conclusions in threatening environments may be adaptive as it can lead to increased precautionary action.

Keywords: Threat, Anxiety, Stress, Evidence Accumulation, Valence, Sequential Sampling

Significant Statement

To make good judgments people gather information. As information is often unlimited a decision has to be made as to when the data is sufficiently strong to reach a conclusion. Here, we show that this decision is significantly influenced by perceived threat. In particular under threat the rate of negative information accumulation increased, such that weaker evidence was required to reach an undesirable conclusion. Such modulation could be adaptive as it can result in enhanced cautious behavior in dangerous environments.

Many important decisions are made when people feel stressed and anxious (Beilock, 2010). Consider a doctor in the operating theatre who needs to decide on the best course of action, a soldier on the battlefield who must decide whether to attack, or a driver stuck in traffic selecting which route to take. Whether calm or stressed, to make good decisions people need to gather information over time (Platt & Glimcher, 1999; Usher & McClelland, 2001). For example, a doctor may decide to consult multiple colleagues before deciding to amputate. Because information can be unlimited, an agent needs to determine when the available data is strong enough to make a judgement (Gluth et al., 2012, 2013). Here, we examine how perceived threat impacts the process by which evidence is accumulated to reach a judgement.

A feature of threatening environments is that the potential for adverse outcomes is high. In these instances, it is adaptive to err on the side of caution. For example, imagine you are walking through a dark alley and hear a ‘pop’. The sound may be a gunshot or perhaps uncorking of a champagne bottle. Interpreting the sound as the former will cause you to escape and mitigate potential risk. Thus, under perceived threat it may be adaptive to interpret a stimulus as undesirable even if the strength of the evidence supporting such judgement is only limited. The psychophysiological reaction induced by threat can provide a global, rather than specific, danger signal. We thus hypothesized that the effects of threat on evidence accumulation may be observed even when the source of the threat is unrelated to the decision at hand (e.g., a psychophysiological reaction triggered by a professional conflict may impact how the ‘pop’ is interpreted).

Computationally, this process may occur in at least two ways. First, under perceived threat people may be predisposed towards undesirable judgments before attaining any information (e.g., you may believe the road you are walking down is dangerous before observing any evidence to that effect). A second, not mutually exclusive possibility, is that under perceived threat an undesirable piece of evidence (e.g., an anxious looking man walking down the road) drives beliefs towards an undesirable judgment ('this road is dangerous'), more so than a desirable piece of evidence (e.g., people are walking past you relaxed and happy) towards a desirable judgment ('this road is safe'). These two distinct mechanisms will result in the same observable behavior. In particular, weaker evidence will be needed to support undesirable judgments under perceived threat.

To tease apart these mechanisms, we used a sequential sampling model to model noisy evidence accumulation towards either of two judgment thresholds (Ratcliff, 1978; Ratcliff & Rouder, 1998; Voss et al., 2013). The model allows estimation of both (i) starting point and (ii) rate of evidence accumulation, reflecting the quality of information processing. We can then measure whether either of these factors are influenced by the desirability of a judgement and how this is influenced by perceived threat.

We exposed participants to an acute threat manipulation in the lab (Garrett et al., 2018), or a control condition, and then asked them to complete an evidence accumulation task (Gesiarz et al., 2019) that was unrelated to the cause of the threat. In the task, participants witness various stimuli that are contingent upon which one of two hidden states they are in. One state was associated with greater rewards than losses (desirable state) and the other with greater losses than rewards (undesirable state). Participants had no control over which state they were in; their task was simply to judge the state, gaining additional rewards for accurate judgments and losing rewards for inaccurate judgments. Thus, it is in participants' interest to be as accurate as possible. They were allowed to accumulate as much evidence as they wished before making a judgment. We examine if and how perceived threat impacts the accumulation of evidence towards a judgement.

Methods.

Experimental Design.

Participants. A total of 91 individuals participated in this study at two sites: University College London (UCL, N = 51) and Massachusetts Institute of Technology (MIT, N = 40). They were recruited via the participant pools of UCL and MIT. All analyses were repeated separately for participants tested in the two different locations (MIT, UCL). There were no differences between locations in any of our results including model-free analysis, psychometric equations or DDM analysis.

Participants gave written, informed consent and were remunerated £7.50/$15 for their participation plus an unspecified performance-related bonus. Ethical approval was provided by the Research Ethics Committees at UCL and MIT. One participant who terminated the experiment early and another who failed all comprehension checks were excluded from the analysis. In addition, we followed the exclusion criteria previously published for this task

(Gesiarz et al., 2019): we excluded two participants whose accuracy rate was below chance (50%) and four who provided responses based only on the first stimulus in over half the trials. Thus, data of 83 participants was included in the analysis (M_{age} = 30.29., SD_{Age} = 12.20; 37 females, 46 males, 43 at UCL and 40 at MIT). Each participant was randomly assigned to either the threat manipulation group (N = 40, M_{age} = 28.98, SD_{Age} = 11; 14 females, 26 males, 21 at UCL and 19 at MIT) or the control group (N = 43, M_{age} = 31.51, SD_{Age} = 13.23; 23 females, 20 males, 22 at UCL and 21 at MIT).

Manipulation procedure and Manipulation Check. We followed the exact same threat manipulation as in Garrett et al. (2018). Participants assigned to the threat manipulation group were informed that at the end of the experiment they would be required to deliver a speech on a surprise topic, which would be recorded on video and judged live by a panel of staff members. They were shown an adjacent room where chairs and tables were already organized for the panel. This manipulation is a variation of the Trier Social Stress Test (TSST; Birkett, 2011) with the key difference being that participants in this task were threatened by the possibility of a stressful social event and completed the main task under anticipation of the threat, but the threat was never executed. Having the participants believe the threatening event will take place at the end of the task, rather than before, increased the likelihood that participants' anxiety levels remained high throughout the task. In addition, participants were presented with six difficult mathematical problems that they were asked to try and solve in 30 s. The exact same manipulation procedure previously executed in our lab, has been shown to significantly heighten cortisol levels, skin conductance and self-reported state anxiety (Garrett et al., 2018).

We have also shown that the manipulation-induced changes in self-reported state anxiety (measured using the Spielberger State Trait Anxiety Inventory) correlated across participants with physiological indicators of stress (Garrett et al., 2018).

Participants assigned to the control group were informed that at the end of the experiment they would be required to write a short essay on a surprise topic, which would not be judged. They were then presented with six elementary mathematical problems to solve in 30 s. This control manipulation has been shown not to heighten cortisol levels, skin conductance and self-reported state anxiety (Garrett et al., 2018). As a manipulation check, before and after the induction procedure, we asked participants to complete the Spielberger State Trait Anxiety Inventory (STAI, Marteau & Bekker, 1992) as a measure of anxiety.

Behavioral Task. After completing the threat/control manipulation, participants played 80 trials of the "Factory Game" , published previously by Gesiarz et al., (2019). On each trial participants witnessed an animated sequence of televisions and telephones passing along a conveyor belt. There were two types of trials: Telephone Factory trials and Television Factory trials. In telephone factory trials the probability of each item in the animated sequence being a telephone was 0.6 and of being a television 0.4. For Television Factory trials the proportions were reversed. The trial type was randomly determined with replacement on every trial with an equal probability for each trial type. Participants were tasked with judging whether they were in a Telephone Factory trial or a Television Factory trial. Since the trial type was not directly observable, their means of doing this was through reverse inference over the sequence of objects they were seeing. Participants were free to respond as soon as they wished after initiating the trial and the sequence would continue until they made their choice.

Participants began the game with an endowment of 5000 points. Each 100 points was worth 1 pence/1cent. One of the two factory types was randomly assigned per participant to be the desirable factory type and the other to be an undesirable type. Participants were informed that each time they visited the desirable factory (desirable state), they would win points, and each time they visited the undesirable factory (undesirable state), they would lose points. We did not specify the exact number of points they will win or lose. Crucially, this bonus was entirely outside of the participants' control, i.e., it was not affected by the judgments the participants made. Separately, participants were informed that they would earn an unspecified number of points for making a correct judgment and lose an unspecified number of points for making an incorrect judgment. We informed subjects that the magnitude of each unspecified bonus/loss were independent of each other, potentially unequal and varied randomly on each trial.

The task was the same as published previously (Gesiarz et al., 2019), except that we jittered the presentation time of the stimuli, so that participants were less likely to have a clear expectation of when the next stimulus would be observed. Due to a technical error this jitter was slightly different across sites (average stimuli presentation time at UCL: 657.28ms, SD = 1060.73ms; MIT: 373.65ms, SD = 49.69ms). The lag between stimuli was ~150 ms.

Trials in which participants made their judgment before observing the second object were removed. In cases where a participant did this in over half their trials, we assumed that the participant was not appropriately engaging with the task and eliminated the entirety of their trials. Following Gesiarz et al., (2019) we dropped four participants for this reason, as well as a further 72 responses made before seeing the second item.

Training. Prior to playing the task, participants received extensive instructions and were required to answer multiple-choice comprehension check questions on the key points of the task, with the question repeated until they either chose correctly or failed three times, upon which the correct answer was displayed. The comprehension check questions addressed the following key points of how the game worked: that telephone factories mostly produced telephones, but sometimes produced televisions; the bonus for visiting desirable factories was independent of the judgments they made; which factory was their desirable factory; and that trial types (i.e., if they were in a TV or phone factory) were randomly determined and it was not guaranteed that they would see exactly the same amount of each type of factory. Participants then played a practice session of 20 trials, where the trial type was visibly displayed to them (i.e., if they were in a TV or phone factory), so they could have prior experience of the outcome contingencies and the trial type distribution.

[insert Figure 1 here]

Statistical Analysis.

Manipulation Check. An independent two-tailed t-test was computed to assess the difference in percentage change in STAI ((post STAI- pre STAI)/pre STAI) between the threat and control group. One-sample t-tests were computed to assess percentage change against zero within each group.

Psychometric Function. We followed the same analysis as in Gesiarz et al., (2019) to relate participants' judgments to the strength of evidence they observed. We fitted a psychometric function, using a generalized mixed effects equivalent of a logistic regression, with fixed and random effects for all independent variables. We fitted these functions separately for participants for whom TV factory was desirable and for whom TV factory was undesirable, and separately for each group (control, threat).

$$P(TV) = \frac{1}{1 + e^{-(\beta_1 X - \beta_0)}}$$

Where P(TV) is the probability of a participant indicating they are in a TV factory; X is the proportion of TV stimuli out of all stimuli observed in a trial. This variable was centered, thus ranging from 0.5 when all samples were TVs to -0.5 when all samples were phones; β_0 is the indifference point – reflecting the proportion of TVs required to respond TV 50% of the time. If $\beta_0 = 0$, participants would indicate they are in a TV factory half the time when half the samples were TVs. When β_0 is low the function will move left and vice versa; β1 is the slope, reflecting by how much the probability of a participant indicating they are in a TV factory increases when the proportion of TVs increases by one unit.

Drift-diffusion modelling. Our aim in modelling our task using the drift-diffusion framework was to assess how perceived threat impacted the evidence accumulation process. In particular, we wanted to assess 1) whether the evidence accumulation process in the threat and control groups was best represented by the same model or a different model, and 2) whether perceived threat impacted the parameters of the evidence accumulation process in our data.

We implemented and compared four different specifications of a DDM (**see Table 1**). The models included the following parameters: (1) t0—amount of non-accumulation/non-decision time; (2) α —distance between decision thresholds; (3) z—starting point of the accumulation process; and (4) v–drift rate - is the rate of evidence accumulation. Crucially, in models 1 and 3 the starting point was fixed to 0.5, while in models 2, 4 we allowed the starting point to vary towards one threshold (its value could vary between 0 and 1, thus allowing a valence-dependent starting point bias). In models 1,2 with an unbiased drift rate, the parameter was symmetric for desirable and undesirable factories (v and -v). In models 3,4 we allowed the drift rate to vary (which we call a valence-dependent drift rate bias) depending upon whether the participant was visiting a desirable factory or an undesirable factory (thus allowing a process bias). In these models we included a term reflecting the difference between drift rates for desirable and undesirable factories (β1factory desirability). "Factory desirability"—is the true factory visited coded as 1 for desirable factories and 0 for undesirable factories. Positive values indicated a bias towards desirable judgements, and negative values indicated a bias towards undesirable judgements. β0 is a constant for the drift rate.

[insert Table 1 here]

We used the HDDM software toolbox (Wiecki, Sofer, & Frank, 2013) to estimate the parameters of our models. The HDDM package employs hierarchical Bayesian parameter estimation, using Markov chain Monte Carlo (MCMC) methods to sample the posterior probability density distributions for the estimated parameter values. We estimated both group-level parameters as well as parameters for each individual participant. Parameters for individual participants were assumed to be randomly drawn from a group-level distribution. Participants' parameters both contributed to and were constrained by the estimates of group-level parameters. In fitting the models, we used priors that assigned equal probability to all possible values of the parameters. Models were fit to log-transformed RTs as done previously (Gesiarz et al., 2019), because RTs were non-normally distributed and had a heavy positive skew. Also, since our "error" RT distribution included relatively fast errors we included an inter-trial starting point parameter (sz) for both models to improve model fit (Ratcliff & Rouder, 1998). We sampled 20000 times from the posteriors, discarding the first 5000 as burn in and thinning set at 5. MCMC are guaranteed to reliably approximate the target posterior density as the number of samples approaches infinity. To test if the MCMC converged within the allotted time, we used Gelman-Rubin statistic (Rubin & Gelman, 1992) on 5 chains of our sampling procedure. The Gelman–Rubin diagnostic evaluates MCMC convergence by analyzing the difference between multiple Markov chains. The convergence is assessed by comparing the estimated between-chains and within-chain variances for each model parameter. In each case, the Gelman-Rubin statistic was close to one (<1.1), suggesting that MCMC were able to converge. To assess if the parameters describing the bias in prior and drift rate are significantly different in the control and threat group, we compared 95% confidence intervals of the parameters' values. In addition, model fits were compared using the Deviance information criterion (Spiegelhalter et al., 2002), which is a generalization of the Akaike Information Criterion (AIC) for hierarchical models. The DIC is commonly used when the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation (Gamerman & Lopes, 2006). It allows one to assess the goodness of fit, while penalizing for model complexity.

To validate the winning model, we used each group's parameters obtained from participants' data to simulate log RTs and judgments separately for the threat and control group. We used the exact number of subjects, total number of trials and trial structure as in the experiment. Simulated data was then used to (i) perform model recovery analysis and (ii) to compare the pattern of participants' response to the pattern of simulated responses, separately for each group. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. Simulation and model recovery analysis were performed using the HDDM software toolbox (Wiecki et al., 2013).

Proportion of correctly identified factories. We computed a linear mixed effects model to assess how group (control/threat) and valence of factory visited (desirable factory/undesirable factory) affected proportion of correctly identified factories as desirable or undesirable. Group, valence of factory, and group*valence of factory interaction were included as fixed and random variables. We included both fixed and random intercepts. We compared the pattern of results obtained from participants' real data to those obtained from the simulated data.

Results.

Threat manipulation was successful. The manipulation was successful in inducing perceived threat. Participants in the threat group reported a significantly higher increase in anxiety as a result of the manipulation (increase in STAI score after the manipulation relative to before M = 40.82%, SD = 37.49, $t(39) = 6.89$, $p<0.001$), compared to those in the control group, who in fact showed a reduction in anxiety (M = -5.74%, SD = 8.24, $t(42) = -4.57$, $p<0.001$, difference between the two groups: $t(82) = -7.94$, $p<0.001$, d' = 1.715, **Figure 2**) an effect often observed in control participants, who tend to relax as they learn more about the task at hand (Garrett et al., 2018).

[insert Figure 2 here]

Under threat participants required weaker evidence to conclude they are in an undesirable factory. We first examined whether perceived threat alters the strength of evidence participants require to make desirable and undesirable judgements. To that end, we fit a psychometric function to the data which relates the percentage of TVs observed on a trial (i.e., the strength of the evidence to judge a factory as TV) to participants' judgment on whether they are visiting a TV or telephone factory. This was done separately for participants for whom the TV factory was desirable and for whom it was undesirable in the threat and control group.

As observed in **Figure 3a,** under perceived threat the psychometric function of participants for whom the TV factory was undesirable (solid orange line) was shifted left compared to controls (dotted orange line). This means that compared to control, under threat participants required a smaller proportion of TVs to be observed before reaching the conclusion that they were in a TV factory when the TV factory was undesirable (indifference parameter was higher for the threat group: $\beta 0$ = 0.11, 95% CI [-0.19, 0.41], than controls: $\beta 0$ = -0.58, 95% CI [-0.96, -0.20], d' = 0.61, **Figure 3a**). No such difference is observed when the TV factory is desirable; participants in both groups require an equal proportion of TVs to be observed before reaching the conclusion that they were in a TV factory. This can be seen in **Figure 3b** where the psychometric function for threat and control participants overlap (indifference parameter was not different for the threat group: $\beta 0$ = 0.28, 95% CI [-0.07, 0.62] and control group: $\beta 0$ = 0.23, 95% CI [-0.08, 0.53], d' = 0.045, **Figure 3b**).

While controls required weaker evidence to conclude they were in a desirable factory than undesirable factory (replicating previous findings - Gesiarz et al., 2019), the difference was abolished under perceived threat. This can be observed where the psychometric function of control participants for whom the TV factory was desirable (dotted blue line, **Figure 3c**) is shifted to the left of control participants for whom the TV factory was undesirable (dotted orange line, **Figure 3c**) (indifference parameter was greater for desirable factory: $\beta 0$ = 0.23, 95% CI [-0.08, 0.53] than undesirable: $\beta 0$ = -0.58, 95% CI [-0.96, -0.20], d' = 0.694, **Figure 3c**), while for participants in the threat group they overlap (indifference parameter when the TV factory was desirable $\beta 0 = 0.28$, 95% CI [-0.07, 0.62] and undesirable $\beta 0 = 0.11$, 95% CI [-0.19, 0.41], $d' = 0.157$, **Figure 3d**).

As expected, both in the threat and control group the greater the proportion of TVs in a trial the more likely participants were to judge the factory as a TV factory (control: TV factory desirable: $\beta 1 = 25.55$, 95% CI [23.20, 27.90], TV factory undesirable: $\beta 1 = 24.94$ [15.62, 34.26], $d' = 0.026$, **Figure 3c**; threat: TV factory desirable: $\beta 1 = 27.79$, 95% CI [19.17, 36.41], TV factory undesirable: $\beta 1 = 23.04$, 95% CI [17.31,28.78], $d' = 0.201$, **Figure 3d**).

[insert Figure 3 here]

Note, that the total number of pieces of evidence (televisions + telephones) did not differ when participants reached an undesirable or desirable conclusion ($F(1,82.77) = 1.39$, $p = 0.24$), nor did it differ as a function of perceived threat ($F(1,81.95) = 1.24$, $p = 0.27$), neither was there an interaction between these two factors ($F(1,82.77) = 1.66$, $p = 0.20$). Rather, as shown above, it is the proportion of evidence (which signifies the strength of the evidence) needed to reach a conclusion that differed as a function of perceived threat and valence.

Thus far our analysis suggests that perceived threat leads to a reduction in the strength of the evidence needed to reach undesirable conclusions, even though the cause of the threat (anticipating a negative social situation) had nothing to do with the task at hand. We next sought to identify the precise computational factor affected by perceived threat during evidence accumulation.

Under threat the drift rate towards undesirable conclusions is greater. Computationally, there are at least two different ways by which perceived threat can lower the strength of evidence needed to reach undesirable conclusions. First, threat may alter the starting point of the accumulation process. That is, if under perceived threat participants are a-priori more likely to believe they are in an undesirable state relative to controls than weaker evidence will be needed to reach that conclusion. Alternatively, perceived threat can enhance the weight given to each piece of negative evidence relative to control. This again will lead to weaker evidence needed to reach an undesirable conclusion.

To tease apart these possibilities we modelled the responses as a drift-diffusion process (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013) with the following parameters: (1) t0—amount of non-accumulation/non-decision time; (2) α —distance between decision thresholds; (3) z—starting point of the accumulation process; and (4) v–drift rate - is the rate of evidence accumulation (**for details see Methods**). Crucially, in models 1 and 3 the starting point was fixed to 0.5, while in models 2, 4 we allowed the starting point to vary towards one threshold (thus allowing a starting point bias). In models 3,4 we allowed the drift rate to vary (which we call a drift rate bias) depending upon whether the participant was visiting a desirable factory or an undesirable factory (thus allowing a process bias).

The Deviance Information Criterion (DIC), a generalization of the Akaike Information Criterion for hierarchical models, was calculated for each model **(Table 2)**. The DIC scores indicated that Model 4 (the valence-dependent model), which included a valence-dependent starting point and drift rate, outperformed all other models for both threat and control groups. As can be observed in **Figure 4,** while for the control group the valence-dependent model was clearly a better fit than the valence-independent model (replicating our previous results Gesiarz et al., 2019), for the threat group the advantage in terms of fit was modest.

[insert Table 2 here]

[insert Figure 4 here]

We next examined which of the accumulation parameters were affected by perceived threat. As observed in **Table 3** and **Figure 5**, only one element in the accumulation process was significantly altered by perceived threat: the valence-dependent drift rate bias. The drift rate bias is the difference in drift rates between desirable and undesirable factories, the greater the bias the greater the drift rate for desirable factories relative to undesirable ones. As can be observed in **Figure 5e** the valence-dependent bias in drift rate in the control group was significantly greater than in the threat group (control: $\beta 1 = 0.17$ [0.07, 0.27]; threat: $\beta 1 = -0.08$ [-0.20, 0.04]). For controls the bias in drift was significantly *positive* (95% confidence intervals (CI) do not include zero: $\beta 1 = 0.17$ [0.07, 0.27]), leading to a drift rate that was more than double when participants were in the desirable factory ($v_{desirable} = 0.63$) than undesirable factory

($v_{undesirable}$ = 0.46). In contrast, under perceived threat the bias in drift rate was numerically negative and not significantly different from zero (95% confidence intervals (CI) include zero: β1 = -0.08 [-0.20, 0.04]), leading to a drift rate that was numerically and non-significantly larger when participants were in the undesirable factory ($v_{undesirable}$ = 0.63) than desirable factory ($v_{desirable}$ = 0.55).

[insert Table 3 here]

[insert Figure 5 here]

We simulated data using group parameters from the threat and control group separately (**for details see Methods**). We first examined if the model parameters could be successfully recovered based on the simulated data. To do so the valence-dependent model was fit to simulated data, in the same way as for the experimental data. We sampled 2000 times from the posteriors, discarding the first 500 as burn in. As shown in **Table 4** model parameters could be successfully recovered based on the simulated data. Additionally, we examined if that the simulated data reproduced the same behavioral pattern of results as the participants' data. This was indeed the case (see **Figure 6d,** detailed explanation below).

[insert Figure 6 here]

[insert Table 4 here]

As DDM parameters are computed partially based on participants' judgements we expected the model-based valence-dependent drift rate bias to correlate across individuals with a valence-dependent bias in judgements. Indeed, across participants there was a strong positive correlation between valence-dependent drift rate bias and the proportion of correctly identified desirable factories minus the proportion of correctly identified undesirable factories (threat group: $r = 0.802$, $p<0.001$, **Figure 6a,** control: $r = 0.918$, $p<0.001$, **Figure 6a),** which we term 'valence-dependent judgement bias'. Individuals with greater drift rate towards desirable than undesirable judgements were more likely to correctly identify desirable factories as desirable when they observed them than undesirable factories when they observed them. In contrast, starting point bias did not correlate with a valence-dependent bias in judgements in the threat group ($r = 0.223$, $p = 0.191$), but did in the control group ($r = 0.517$, $p<0.001$). In the latter, a larger starting point bias was related to the proportion of correctly identified desirable factories minus the proportion of correctly identified undesirable factories (**Figure 6b**).

As we have already shown that participants in the control group had a greater drift rate bias than those under perceived threat, it follows that they will also show greater valence-dependent judgement bias. This is exactly what we found. Entering whether a judgement was correct (coded as 1 for correct response and 0 for incorrect) on every trial into a mixed linear model with valence of factory, group and their interaction as fixed and random effects, as well as fixed and random intercepts revealed a group by valence interaction ($F(1,77.46) = 4.67$, $p =$

0.03, **Figure 6c**) as well as a main effect of factory valence ($F(1,77.46) = 5.96$, $p = 0.02$) and no main effect of group ($F(1,80.77) = 0.02$, $p = 0.90$). To tease apart the interaction we ran the same linear mixed models as above separately for each group. This revealed an effect of valence in the control group ($F(1,77.87) = 11.06$, $p = 0.001$), where participants were less likely to correctly categorize undesirable factories (proportion of correctly categorized factories undesirable factories $M = 0.72\%$, 95% CI [0.68, 0.76]) than desirable factories (proportion of correctly categorized factories desirable factories: $M = 0.80\%$, 95% CI [0.76, 0.84]. In contrast, under perceived threat the effect of valence dissappeared ($F(1,80.77) = 0.02$, $p = 0.90$); participants were as likely to correctly categorize undesirable factories ($M = 0.77\%$, 95% CI [0.73,0.80]) as they were desirable factories ($M = 0.76\%$, 95% CI [0.72, 0.80]). This suggest that under perceived threat the valence-dependent judgement bias is abolished.

We conducted the same analysis on our simulated data and find that it nicely reproduced the behavioral pattern of results (**Figure 6d**).

Discussion

The findings show that perceived threat has a profound effect on the process by which evidence is accumulated. In particular, it leads to a reduction in the strength of the evidence needed to reach undesirable conclusions. Relative to controls, participants under perceived threat required a smaller proportion of negative stimuli to be observed before reaching an undesirable judgement. In contrast, there was no difference between the groups in the strength of evidence accumulated before reaching a desirable judgement. We found this to be true despite the fact that the cause of the threat (anticipating a socially stressful event) was unrelated to the task performed (judging whether more phones or more TVs were observed).

Computationally, there are different mechanisms by which perceived threat can lower the strength of evidence needed to reach undesirable judgements. First, under threat participants may be a priori more likely to believe they are in an undesirable state relative to controls leading to weaker evidence needed to reach that conclusion. Another possibility is that perceived threat can selectively increase the rate of negative information accumulation (drift rate) relative to control. This again will lead to weaker evidence required to reach an undesirable judgement. To tease apart these possibilities we modelled the responses as a drift-diffusion process (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013). We found support for the latter. Specifically, perceived threat altered only one feature of the accumulation process: the relative drift rate towards desirable and undesirable judgments (the 'valence-dependent drift rate bias'). For controls the bias in drift rate was significantly positive – the rate of information accumulation was greater towards desirable than undesirable conclusions (as observed before Gesiarz et al., 2019). Under threat, however, the bias disappeared due to the drift rate towards undesirable judgement increasing.

The results fit with previous suggestions that perceived threat directs attention towards negative stimuli (Riaz et al., 2017) and leads to greater impact of such stimuli on belief updating (Garrett et al., 2018). Indeed, it is possible that the effect of perceived threat on the rate of negative information accumulation is partially due to increased attention towards negative stimuli. The current findings go beyond these previous demonstrations to illuminate the effects of perceived threat on the process of sequential accumulation and show that weaker evidence is needed to reach undesirable conclusions under threat.

Here, we show a causal link between perceived threat and evidence accumulation in healthy individuals. It is interesting, however, to consider how the findings may be related to evidence accumulation in individuals with affective disorders, as these are often triggered by stressful events and/or characterized by high anxiety. With regards to individuals with high trait anxiety, a processing advantage for threatening words has been previously reported (White et al., 2010). While that study was correlational and thus could not determine whether anxiety caused the changes to drift rate and/or vice versa, our results support the notion that anxiety can in fact alter the drift rate towards undesirable conclusions, even if the anxiety is short lived rather than chronic. With regards to individuals with anxiety and mood disorders, one study (Aylward, et al., 2019) found a lower drift rate towards desirable conclusions compared to healthy individuals. Interestingly, the latter study did not detect any effects of induced threat, which may be due to the fact that the task used in that study (as well as in all the above-mentioned studies) unlike ours, was a non-sequential perceptual decision-making task. The process by which pieces of evidence are accumulated over time may be especially impacted by perceived threat.

Our study suggests that evidence accumulation is a flexible process which quickly adjusts to the environment. In particular, the findings show that perceived threat leads to a valence-dependent change to the accumulation process, even when information is not directly related to the cause of the threat. An increased rate of negative information accumulation can then enhance the probability of taking precautionary action to avoid aversive consequences. As aversive outcomes can be more severe and frequent in threatening environments, such generalization can be, on average, adaptive. However, in individuals who are hypersensitive to threat and/or falsely perceive situations as threatening, such as those suffering from anxiety and depression, an increased rate of negative information accumulation could be maladaptive. This is because such increased rate can produce overly pessimistic predictions, which induce stress and anxiety further, elevating symptoms.

References.

Aylward, J., Hales, C., Robinson, E., & Robinson, O. J. (2019). Translating a rodent measure of negative bias into humans: The impact of induced anxiety and unmedicated mood and anxiety disorders. *Psychological Medicine*, *50*(2), 237–246. https://doi.org/10.1017/S0033291718004117

Beilock, S. (2010). *Choke: What the secrets of the brain reveal about getting it right when you have to*. Simon and Schuster.

Birkett, M. A. (2011). The Trier Social Stress Test protocol for inducing psychological stress. *Journal of Visualized Experiments*, (56), 1–6. https://doi.org/10.3791/3238

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *Journal of Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Gesiarz, F., Cahill, D., & Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLoS Computational Biology*, *15*(6), e1007089. https://doi.org/10.1371/journal.pcbi.1007089

Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*(31), 10686–10698. https://doi.org/10.1523/JNEUROSCI.0727-12.2012

Gluth, S., Rieskamp, J., & Büchel, C. (2013). Deciding Not to Decide: Computational and Neural

Evidence for Hidden Behavior in Sequential Choice. *PLoS Computational Biology*, *9*(10). https://doi.org/10.1371/journal.pcbi.1003309

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology*, *31*(3), 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. https://doi.org/10.1038/22268

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). Drift Diffusion Decision Model:Theory and data. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Riaz, N., Wolden, S. L., Gelblum, D. Y., & Eric, J. (2017). Attention Bias towards Negative Emotional Information and its Relationship with Daily Worry in the Context of Acute Stress: An Eye-Tracking Study. *Behaviour Research and Therapy*, *90*(24), 96–110. https://doi.org/10.1002/cncr.27633.Percutaneous

Rubin, D. B., & Gelman, A. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. Retrieved from http://www.jstor.org/stable/2246093%0Ahttp://dash.harvard.edu/handle/1/3630270

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical*

Methodology, *64*(4), 583–616. https://doi.org/10.1111/1467-9868.00353

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, Vol. 108, pp. 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Anxiety Enhances Threat Processing Without Competition Among Multiple Inputs: A Diffusion Model Analysis. *Emotion*, *10*(5), 662–677. https://doi.org/10.1037/a0019474

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*(JULY 2013), 1–10. https://doi.org/10.3389/fninf.2013.00014

Figure Legends.

Figure 1. The factory task. In each trial, participants see an animated sequence of televisions and telephones passing along a conveyor belt. Their task is to accurately determine whether they were in a) a telephone factory, i.e., a factory that produces telephones most of the time or b) a television factory, i.e., a factory that produces televisions most of the time. They are incentivized for accuracy and can enter their judgment whenever they like. Each participant is “invested” in one factory. On trials where they happen to be in that (desirable) factory they gain points, on trials in which they happen to be in the other (undesirable) factory they lose points. Notably, this bonus is beyond participants’ control and is not affected by the actual judgment made. Stimulus presentation time was jittered, so that participants were less likely to have a clear expectation of when the next stimulus would be observed. Stimulus presentation time on average was ~ 521 ms. The lag between stimuli was on average ~150 ms.

Figure 2. Threat manipulation was success. Participants in the threat group became significantly more anxious after the manipulation than in the control group. Data are plotted as box plots for each condition, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 x interquartile range. Red diamond shape indicates the mean percentage change in STAI per experimental group. Individuals’ percentage STAI change are shown separately as grey circles. *** $p<0.001$

Figure 3. Under Threat weaker evidence is required to reach undesirable conclusions. Fitted psychometric function for data of the threat group (solid line) and control group (dotted line). Y axis shows the proportion of times participants indicated they were in a TV factory as a function of the proportion of TV items they observed in a trial prior to making a judgment (x axis). In blue is the data of participants for whom the TV factory was the desirable factory. In orange is the data of participants for whom the TV factory was the undesirable factory. **(a)** Compared to control, under perceived threat participants required a smaller

proportion of TVs to be observed before reaching the conclusion that they were in a TV factory, when the TV factory was undesirable. This can be seen as the solid line (threat group) is shifted left relative to the dotted line (control group). **(b)** No such difference is observed when the TV factory is desirable. **(c)** Participants in the control group required a smaller proportion of TVs to be observed before reaching the conclusion that they were in a TV factory, when the TV factory was desirable than undesirable. This is seen as the blue line (desirable) is shifted left relative to the orange line (undesirable). **(d)** This difference is abolished under perceived threat. Error bars show SEM at given level of proportion of TVs observed (error bars for threat group are indicated by ‘x’ at the center of the error bar). Grey dashed line indicates point of indifference – i.e., how much evidence is needed for participants to say ‘TV’ half the time.

Figure 4. Difference in fit between winning valence-dependent model and valence-independent model as a function of perceived threat. The Y axis shows the difference in Deviance Information Criterion (DIC) scores between the valence-independent model and the winning valence-dependent models for the control group (dark grey) and threat group (light grey).

Figure 5. Under threat valence-dependent drift rate bias is abolished. Displayed are the posterior distributions of parameter estimates for the threat group (light grey) and the control group (black). Light green line indicates 95% CI for the threat group. Dark green line indicates 95% CI for the control group. No significant difference is observed between groups for estimates of **(a)** decision threshold **(b)** non-decision time **(c)** starting point **(d)** drift-rate constant. **(e)** In contrast, a significant difference is observed for the valence-dependent drift rate bias. In the control group the bias indicates a significantly larger drift rate towards the desirable than undesirable conclusion. This bias is corrected for under perceived threat and is numerically inverse (that is the non-significant bias is negative under perceived threat but significantly positive for controls). * indicates significant difference between parameters in threat and control group (i.e., confidence intervals do not overlap).

Figure 6. Threat-induced change in valence-dependent drift rate bias is expressed as valence-dependent changes in proportion of correctly identified factories. (a) A positive relationship is observed between valence-dependent drift rate bias (Y axis) and valence-dependent bias in proportion of correctly identified factories (X axis). Individuals with greater drift rate towards desirable than undesirable conclusions are more likely to correctly categorize desirable than undesirable factories. This is true both for controls (dark grey) and participants under perceived threat (light grey). For controls the regression line is above that of the participants in the threat group, which is due to the fact that their drift rate bias is significantly greater. The regression line for controls is also shifted to the right which indicates significantly greater valence-dependent judgement bias. **(b)** By contrast we did not observe a relationship between starting point bias (Y axis) and valence-dependent bias in proportion of correctly identified factories (X axis) in the threat group (light grey). A positive correlation was observed in the control group (dark grey). In the control group individuals with a large starting point bias were more likely to correctly identify desirable than undesirable factories. While the line for the threat group is above that of the control this not a significant difference is. **(c)** Controls are less likely to correctly categorize undesirable factories (orange) than desirable factories (blue), while this is not the case for participants in the threat group. **(d)** Simulated Data based on model parameters reproduced these findings. Data are plotted as box plots for each condition, in which horizontal lines indicate median values, boxes indicate 25/75% interquartile range and whiskers indicate 1.5 x interquartile range. Diamond shape indicates the mean. $^{**}p<0.01$, ns = not significant. Clouds represent CIs.

Tables and Legends.

Number	Model	Starting Point (z)	Drift Rate (v)
1.	Valence independent	z = 0.5	v
2.	Valence-dependent starting point	0<z<1	v
3.	Valence-dependent drift rate	z = 0.5	v = β0+β1factorydesirability
4.	Valence-dependent drift rate and starting point	0<z<1	v = β0+β1factorydesirability

Table 1. Drift Diffusion Model Specification. For each group we ran four models which differed in whether we allowed the starting point to vary (model 2, 4), whether we included a valence-dependent drift rate bias (model 3, 4), or neither (model 1).

Number	Model	Starting Point (z)	Drift Rate (v)	DIC (Control)	DIC (Threat)
1.	Valence independent	z = 0.5	v	11373.43	7761.38
2.	Valence-	0<z<1	v	11343.42	7757.88

	dependent starting point				
3.	Valence-dependent drift rate	z = 0.5	v = β0+β1factorydesirability	11322.83	7758.58
4.	Valence-dependent drift rate and starting point	0<z<1	v = β0+β1factorydesirability	11306.45	7744.82

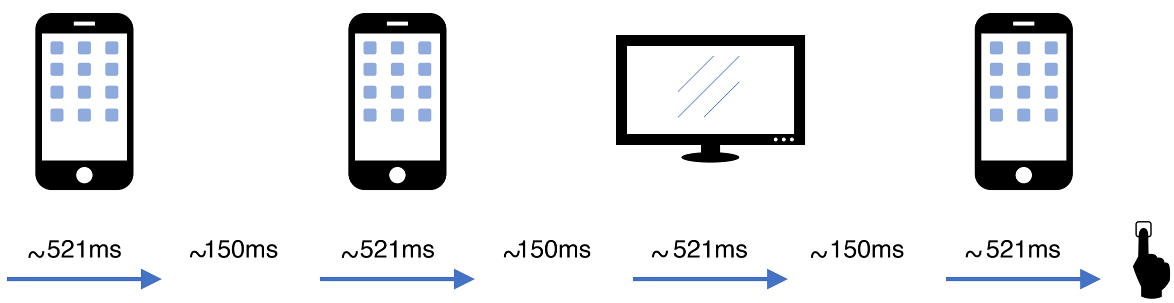
Table 2. Drift Diffusion Model Specification. For each group we ran four models which differed in whether we allowed the starting point to vary (model 2,4), whether we included a valence-dependent drift rate bias (model 3,4), or neither (model 1). DIC scores show goodness of fit, with lower numbers indicating better fit.

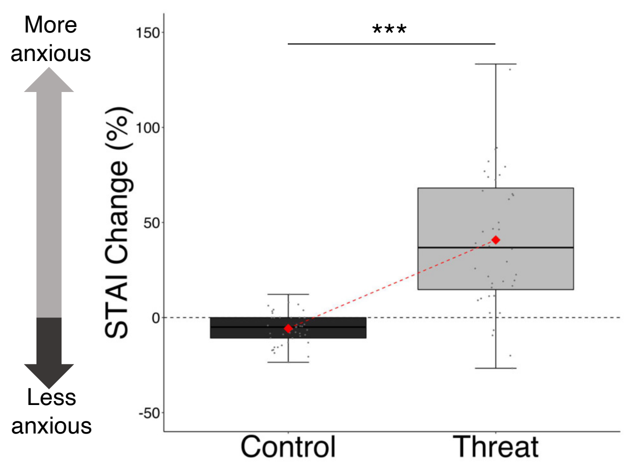
Estimate (from data)	**Control**	**Threat**
Decision Threshold (α)	2.67 [2.49,2.85]	2.47 [2.33, 2.62]
Non-Decision Time (t0)	7.55 [7.37, 7.71]	7.49 [7.33, 7.64]
Starting Point (z)	0.48 [0.47, 0.51]	0.51 [0.50, 0.53]
inter-trial starting point parameter (sz)	0.18 [0.07, 0.27]	0.19 [0.06, 0.28]
Drift Rate (β0)	0.46 [0.37, 0.55]	0.63 [0.54, 0.72]
Drift Rate Bias (β1)	0.17 [0.08, 0.27]	-0.08 [-0.20, 0.04]

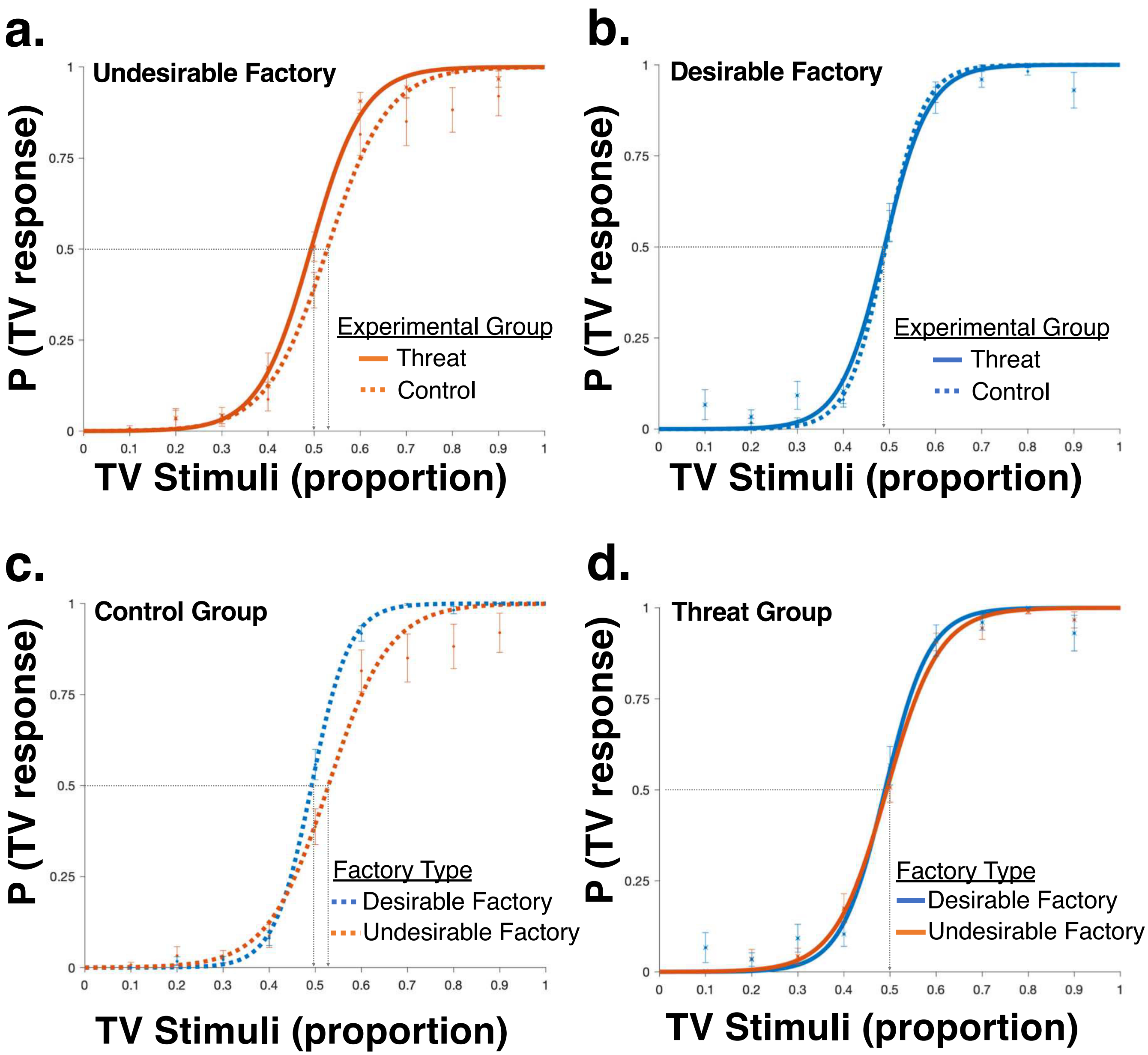
Table 3. Parameter estimates of the evidence accumulation process. Displayed are the model estimates from the winning model for the control and threat groups. These include decision threshold (α), non-decision time (t0), starting point ($0<z<1$), inter-trial starting point parameter (sz), constant drift rate (β0) and drift rate bias (β1). The latter is the term reflecting the additional weight added to the drift rate as a function of factory desirability. Positive values indicate a bias towards desirable judgements, and negative values indicate a bias towards undesirable judgements. [Confidence Intervals].

Estimate (recovered from simulation)	**Control**	**Threat**
Decision Threshold (α)	2.67 [2.63, 2.72]	2.48 [2.43, 2.53]
Non-Decision Time (t0)	7.55 [7.51, 7.59]	7.42 [7.39, 7.46]
Starting Point (z)	0.51 [0.48, 0.53]	0.51 [0.48, 0.54]
inter-trial starting point parameter (sz)	0.33 [0.20, 0.43]	0.15 [0.01, 0.31]
Drift Rate (β0)	0.44 [0.38, 0.51]	0.68 [0.6, 0.77]
Drift Rate Bias (β1)	0.15 [0.08, 0.22]	-0.14 [-0.22, -0.049]

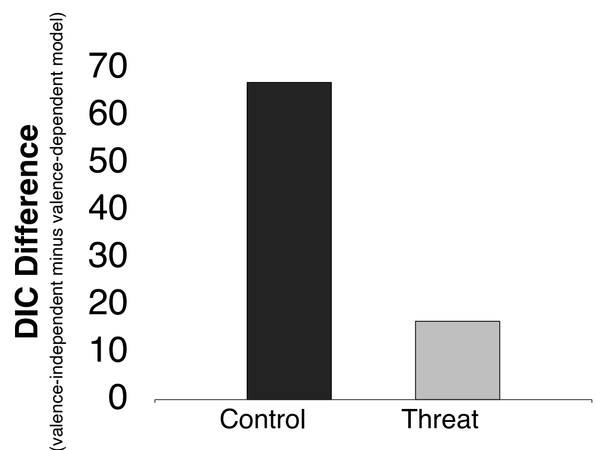
Table 4. Recovered Parameter estimates of the evidence accumulation process based on simulated data. Displayed are the winning model estimates recovered from simulated data for the control and threat groups. These include decision threshold (α), non-decision time (t0), starting point ($0<z<1$), inter-trial starting point parameter (sz), constant drift rate (β0) and drift rate bias (β1). The latter is the term reflecting the additional weight added to the drift rate as a function of factory desirability. Positive values indicate a bias towards desirable judgements, and negative values indicate a bias towards undesirable judgements. [Confidence Intervals].

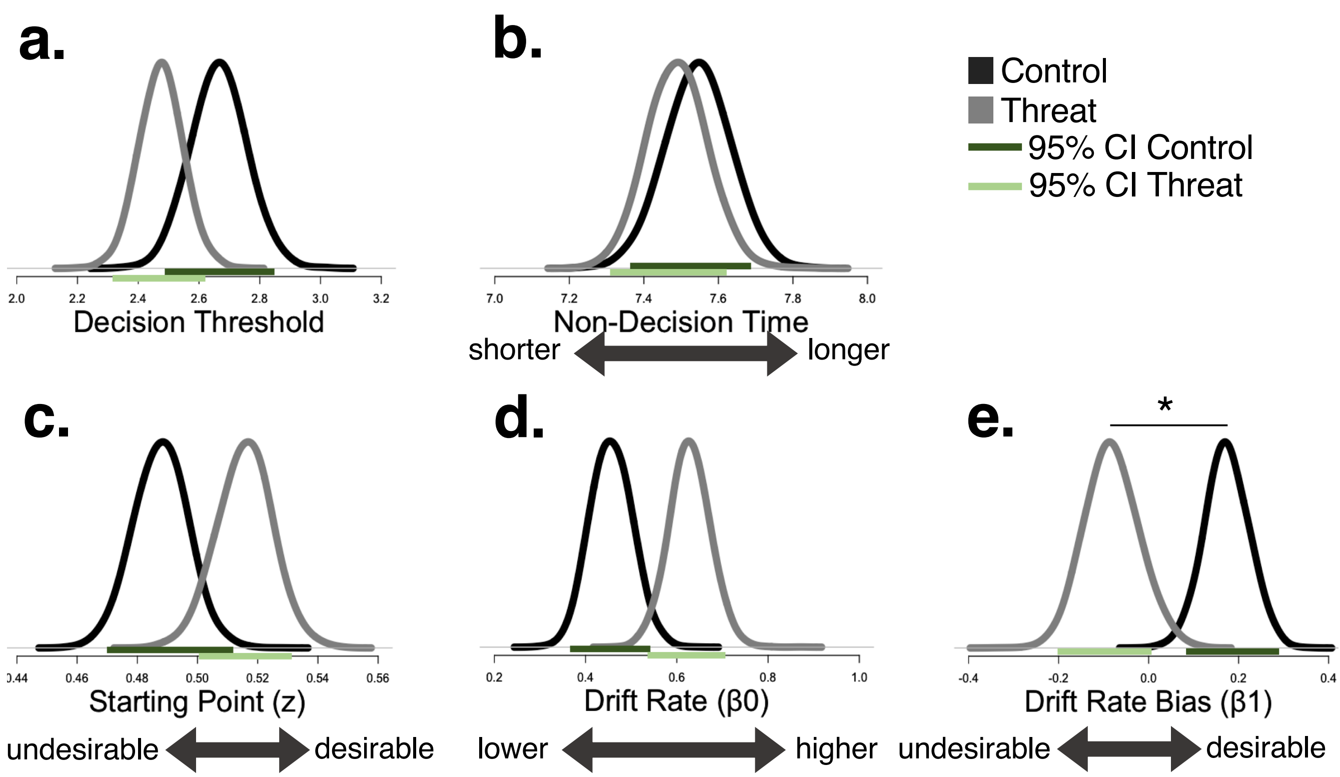

~521ms
~150ms
~521ms
~150ms
~521ms
~150ms
~521ms


More anxious
Less anxious
***
STAI Change (%)
150
100
50
0
-50
Control
Threat

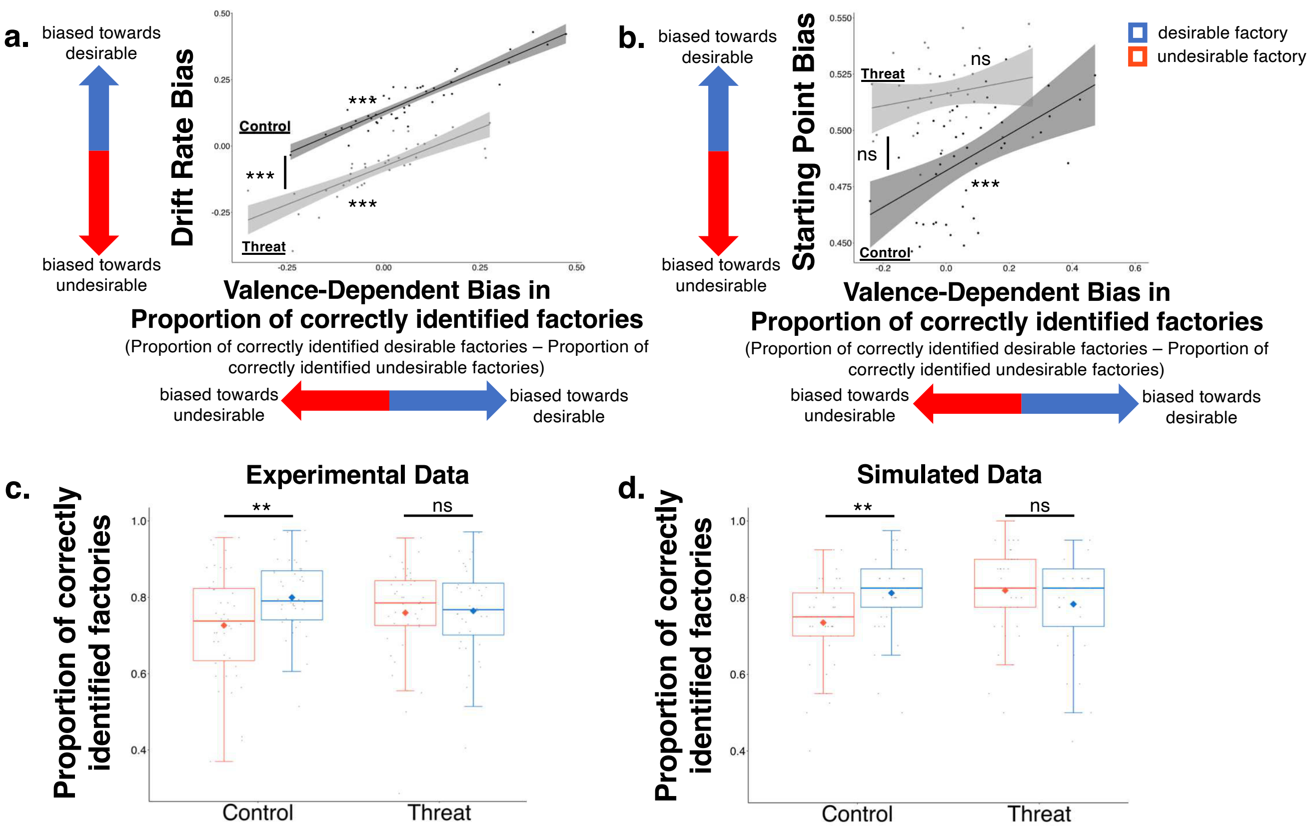


a.
Undesirable Factory
P (TV response)
TV Stimuli (proportion)
Experimental Group
Threat
Control
b.
Desirable Factory
P (TV response)
TV Stimuli (proportion)
Experimental Group
Threat
Control
c.
Control Group
P (TV response)
TV Stimuli (proportion)
Factory Type
Desirable Factory
Undesirable Factory
d.
Threat Group
P (TV response)
TV Stimuli (proportion)
Factory Type
Desirable Factory
Undesirable Factory


DIC Difference
(valence-independent minus valence-dependent model)
70
60
50
40
30
20
10
0
Control
Threat



a.
b.
Control
Threat
95% CI Control
95% CI Threat
2.0
2.2
2.4
2.6
2.8
3.0
3.2
Decision Threshold
7.0
7.2
7.4
7.6
7.8
8.0
Non-Decision Time
shorter
longer
c.
d.
e.
*
0.44
0.46
0.48
0.50
0.52
0.54
0.56
Starting Point (z)
undesirable
desirable
0.2
0.4
0.6
0.8
1.0
Drift Rate (β0)
lower
higher
-0.4
-0.2
0.0
0.2
0.4
Drift Rate Bias (β1)
undesirable
desirable


a.
biased towards desirable
biased towards undesirable
Drift Rate Bias
Control
Threat
***
***
***
Valence-Dependent Bias in
Proportion of correctly identified factories
(Proportion of correctly identified desirable factories – Proportion of correctly identified undesirable factories)
biased towards undesirable
biased towards desirable
b.
biased towards desirable
biased towards undesirable
Starting Point Bias
Threat
Control
ns
ns
***
Valence-Dependent Bias in
Proportion of correctly identified factories
(Proportion of correctly identified desirable factories – Proportion of correctly identified undesirable factories)
biased towards undesirable
biased towards desirable
desirable factory
undesirable factory
c.
Experimental Data
Proportion of correctly identified factories
**
ns
Control
Threat
d.
Simulated Data
Proportion of correctly identified factories
**
ns
Control
Threat