PSYCHOLOGY

Individual differences in naturalistic learning link negative emotionality to the development of anxiety

William J. Villano[1], Noah I. Kraus[1], Travis R. Reneau[2], Brittany A. Jaso[3], A. Ross Otto[4], Aaron S. Heller[1]*

Organisms learn from prediction errors (PEs) to predict the future. Laboratory studies using small financial outcomes find that humans use PEs to update expectations and link individual differences in PE-based learning to internalizing disorders. Because of the low-stakes outcomes in most tasks, it is unclear whether PE learning emerges in naturalistic, high-stakes contexts and whether individual differences in PE learning predict psychopathology risk. Using experience sampling to assess 625 college students' expected exam grades, we found evidence of PE-based learning and a general tendency to discount negative PEs, an "optimism bias." However, individuals with elevated negative emotionality, a personality trait linked to the development of anxiety disorders, displayed a global pessimism and learning differences that impeded accurate expectations and predicted future anxiety symptoms. A sensitivity to PEs combined with an aversion to negative PEs may result in a pessimistic and inaccurate model of the world, leading to anxiety.

INTRODUCTION

Survival and well-being require accurate expectations for the surrounding world. When congruent with reality, expectations for the likely rewards and threats in one's environment prime appropriate behavioral responses and facilitate survival (*1*, *2*), even when information is limited (*3*–*6*). However, as our environments invariably change, expectations that were accurate yesterday might drive us to make the wrong decision today. For example, if a distant rumbling sound elicits an expectation of lightning, one might decide to take cover, even before lightning is visible. Conversely, falsely inferring that the rumble resulted from an airplane might leave the same individual unprepared and surprised when lightning strikes. Surprises such as this, termed prediction errors (PEs), suggest that one's model of the environment is inaccurate (*7*), which may have dire implications for survival (*8*), optimal behavior (*9*), and well-being (*8*).

To maintain accurate expectations in a dynamic world, organisms continuously update their expectations in accordance with recent PEs (*10*–*13*). In the preceding example, a PE results when the occurrence of lightning violates one's expectation of a nearby airplane. Using this PE as a learning signal, an individual may come to expect lightning when they encounter similar rumbling sounds in the future. While the degree to which one changes their expectation typically tracks the size of the PE (*12*–*14*), the influence of PEs on expectation updating varies across people (*15*, *16*) and contexts (*17*, *18*). To account for this variability, reinforcement learning (RL) models [e.g., the Rescorla-Wagner model (*12*)] include a parameter known as the "learning rate," which scales the magnitude of expectation changes relative to PEs (*12*). Critically, learning rates are neither static within nor across individuals (*10*) and vary by several factors including the outcome domain (*19*), environmental volatility (*10*, *20*), contextual familiarity (*21*, *22*), one's learning history (*17*, *23*), and PE valence (*21*, *24*, *25*) (i.e., whether an outcome was better or worse than expected).

Despite the presence of individual differences in learning rates, prior computational work finds that certain regularities in learning rates emerge across a range of RL tasks, such as the tendency to learn differently from positive versus negative PEs (*24*, *26*, *27*). However, the nature of these asymmetries varies considerably across studies, with some results suggesting that individuals are optimistically biased and learn preferentially from positive PEs (*15*, *26*, *28*–*31*), and others finding the opposite pattern (i.e., negativity biases) (*24*, *26*, *32*). Some researchers posit that these opposing asymmetries reflect different cognitive biases, such as loss aversion (*24*) or optimism bias (*33*). However, others suggest that negative valence biases seen in some RL studies (*24*, *26*, *32*) constitute artifacts of specific learning paradigms (*29*) or even model misspecification (*25*). Although most of the evidence supports the presence of an optimism bias in RL (*25*), updating biases have been primarily quantified in low-stakes learning environments with highly concrete, value-based outcomes (e.g., small financial losses and gains, so-called "lower-order" contexts). Thus, it is unclear whether valence-based updating asymmetries are observed in real-world learning contexts where outcomes are often more abstract, pertain to higher-order beliefs, and hold greater personal meaning.

Whereas evidence for optimistic updating in these concrete, value-based RL tasks is mixed, humans are predominantly optimistic when updating abstract beliefs about themselves (*34*–*39*). For instance, when considering abstract quantities, such as one's own ability (*35*, *37*, *40*) or one's attractiveness (*36*), people tend to discount negative feedback, displaying a preference for optimistic albeit potentially inaccurate beliefs (*33*–*36*, *41*). While this tendency to optimistically update higher-order self-relevant beliefs is incompatible with the sole goal of minimizing errors in most RL models [e.g., (*12*, *14*)], optimistic belief updating may present adaptive benefits, such as motivating an individual to persevere despite challenges (*34*, *37*). Moreover, in RL tasks where they emerge, optimistic biases might improve performance beyond that of an unbiased learner (*25*, *42*) by enabling a clearer differentiation between

[1]Department of Psychology, University of Miami, Coral Gables, FL, USA. [2]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA. [3]Center for Anxiety and Related Disorders, Boston University, Boston, MA, USA. [4]Department of Psychology, McGill University, Montreal, Canada. *Corresponding author. Email: aheller@miami.edu

rewarding and punishing options (*42*) and promoting the exploitation of previously rewarded choices (*29*). Thus, a key question is: Why do optimistic biases clearly emerge when humans update higher-order self-relevant beliefs but less so when updating lower-order expectations in RL studies?

One theory is that optimistic updating biases are limited to conditions that are seldom present in RL tasks (*34*). First, optimistic updating biases are typically observed when individuals are faced with feedback regarding beliefs that they care deeply about, commonly termed "motivated beliefs" (*34*, *35*, *37*, *38*). In contrast to most value-based RL tasks with highly concrete outcomes and lower-order expectations (e.g., probability of a win or loss on a gamble) (*24*, *27*, *32*, *43*), motivated beliefs are often ego-relevant beliefs, such as the likelihood that one is more attractive than their peers (*36*, *39*, *41*). Second, optimistic updating biases are more likely when feedback is ambiguous (*34*, *36*). For instance, individuals are quicker to discount negative information when receiving subjective feedback about their attractiveness, relative to objective feedback about their intelligence (*36*), and more so relative to concrete and unambiguous value-based outcomes (*34*). This ambiguous feedback may be more easily discounted, giving rise to the observed asymmetries. Given these two boundary conditions, it may be that when optimistic updating has not been observed in laboratory RL tasks, it is because the outcomes lack personal significance (*34*) or are not sufficiently ambiguous (*24*, *32*).

Another disparity between the value-based RL and motivated belief updating literatures involves the importance of forming accurate expectations. Value-based RL assumes that agents calculate an optimal policy given previous experience (*13*). Influential frameworks of neural processing such as predictive coding posit that minimizing surprise (i.e., PE) via accurate prediction is a core function of the brain (*44*). The importance of accurate prediction is further evident in the brain's error-encoding signals (*45*), which not only drive learning (*46*) but are also aversive (*47*), particularly when outcomes are worse than expected (*48*–*50*). At the same time, a separate literature suggests that individuals often retain objectively inaccurate motivated beliefs in the face of contradicting information (*37*, *39*). While motivated beliefs often manifest as an optimism bias, some individuals display motivated beliefs that are "defensively" pessimistic, whereby individuals overupdate beliefs in a negative direction (*37*, *40*), owing to entrenched core beliefs about their own capacity (*51*). Such a pessimistic updating style lessens the likelihood that one will receive disappointing feedback in the future (i.e., negative PEs) (*52*). Defensively pessimistic biases are observed less frequently than optimistic updating at the population level (*34*) but more commonly in individuals with elevated anxiety (*52*). Together, it may be that motivated reasoning does not only promote optimistic updating generally but also a tendency to disregard certain feedback signals (*36*, *37*, *39*, *40*) in the service of maintaining one's self-beliefs and at the expense of accurate expectations. These beliefs can result in inaccurate expectations due to both optimistic and defensively pessimistic learning biases.

While learning rates vary between and within individuals, a body of recent RL work links variation in learning rates to psychiatric diagnoses (*10*, *20*, *53*–*56*). This work finds that depressed and anxious individuals tend to have larger learning rates for negative PEs, indicating that they learn more from unexpected punishments than rewards. Similarly, other work suggests that optimism biases are attenuated (*59*, *60*) or even supplanted by pessimistic biases in depressed individuals (*16*, *61*, *62*). Moreover, persistently negative expectations, or a very low learning rate to positive PEs, can be an operationalization of defensive pessimism, which is commonly observed in individuals with elevated anxiety (*52*). In contrast, other studies in psychiatric samples highlight general differences in learning rates that are agnostic to PE valence (*15*, *63*). Results from these studies tend to suggest that depressed individuals update expectations more slowly following PEs (*15*), whereas anxious individuals update more rapidly relative to healthy controls (*63*). Given conflicting reports on whether broad variation in learning rates and valence-based learning rate asymmetries are disorder specific, and due to high rates of comorbidity among internalizing disorders (*64*, *65*), it is possible that variation in learning rates constitutes a premorbid risk phenotype for internalizing symptomatology (*18*, *66*).

Although computational psychiatry research has focused primarily on individuals with current internalizing symptoms, a common hypothesis in developmental psychopathology is that elevated sensitivity to negative and unexpected life events (i.e., negative PEs) is a premorbid trait that promotes risk for internalizing disorders given a certain set of life experiences (*67*–*70*). This elevated sensitivity can result in larger learning rates for negative PEs and thus a tendency to overlearn from negative events, leading to pessimistic expectations even in the face of contradictory evidence (i.e., positive PEs). In support of this notion, one study found that adolescents at increased risk for depression are more emotionally reactive to negative PEs (*71*). Moreover, individual differences in personality traits, such as extraversion, consistently predict increased sensitivity to positive PEs in learning tasks (*72*, *73*). While these studies fall short of linking PE sensitivity to differences in PE learning, one recent computational investigation demonstrated that increased sensitivity to negative PEs not only predicted faster learning from negative outcomes but also resulted in inaccurately negative expectations for the future (*74*). Thus, it is plausible that, before the onset of psychopathology, certain personality traits promote individual differences such as increased sensitivity to PEs, which drive the negative learning biases observed in internalizing disorders.

Before the onset of psychiatric disorders, certain personality traits capture temperaments and behaviors that strongly predict the future development of psychopathology (*75*–*80*). To this end, negative emotionality (NE; i.e., neuroticism) is a personality trait, characterized by heightened sensitivity to unexpected and negative events (*68*, *81*, *82*), that emerges early in life (*82*) and predicts the development of anxiety disorders (*75*–*77*). Twin studies suggest that NE accounts for one-third to one-half of genetic liability for internalizing disorders (*80*), and a meta-analysis of prospective studies indicates that a 1-SD increase in NE nearly doubles one's risk for developing an internalizing disorder (*75*). Thus, NE is a risk factor for the development of anxiety (*66*, *77*), and the manner in which individuals learn from PEs may be a mechanism by which NE elevates one's risk (*83*, *84*). Supporting this idea, a body of work suggests that NE increases one's sensitivity to negative life events (*68*) and heightens one's stress responses to common and severe stressors (*68*, *82*). In concert with a learning history of negative life events, heightened stress reactivity in NE can hinder one's ability to discern between threatening and safe stimuli (*68*, *85*–*87*), which, in turn, yields pessimistic expectations for the future. Ultimately, these overgeneralized perceptions of threat, pessimism, and heightened stress reactivity in NE may prompt general distress and behavioral avoidance characteristic of anxiety disorders (*68*, *82*, *83*, *88*, *89*). Thus, given an

accumulation of negative PEs over time, high-NE individuals may overlearn from these negative PEs, ultimately prioritizing the avoidance of such PEs in the future. While one hypothesis is that this type of biased learning may place high-NE individuals at risk for anxiety disorders, no prior work has assessed this possibility.

At present, our understanding of PE learning and its variability across individuals is derived primarily from laboratory-based experimental tasks that differ significantly from real-world contexts (*90*, *91*). These differences limit the conclusions that can be drawn from this work. Typically, laboratory-based tasks investigate PE-driven learning through gambles associated with uncertain probabilities of financial gains or losses. Over hundreds of trials, researchers assume that individuals gradually learn from unexpected losses and gains (PEs), iteratively updating their expectations to eventually converge on accuracy and optimal choice (*43*, *92*). While modeling decision-making over hundreds of trials enables precise estimates of individual learning parameters, this approach has several limitations that hinder our understanding of PE learning.

First, expectations in these laboratory-based tasks are not explicitly sampled but rather inferred by modeling individuals' behavior —a process that requires hundreds of experimental trials per individual. However, in everyday life, a single, personally meaningful PE can have profound effects (*93*–*96*), such as the changes to one's worldview following trauma and subsequent development of posttraumatic stress disorder (*97*). Second, because of the small financial stakes that predominate laboratory-based learning tasks, the consequences of biased learning and resultant inaccurate expectations pale in comparison to real-world, high-stakes contexts in which one's future, well-being, or survival may be at stake. Third, and relatedly, prior work suggests that updating biases are normative when individuals update higher-order, self-relevant beliefs (*34*, *35*, *37*) but are less likely to emerge when updating low-level expectations for inconsequential and highly concrete financial outcomes in RL tasks (*34*). However, there is a paucity of RL studies investigating learning biases with consequential, personally relevant outcomes. Furthermore, extant belief updating studies have failed to explore whether these biased updating styles emerge when feedback is directly experienced and not just informational (*34*); for instance, one belief updating study first asks participants the likelihood they think that they will get a certain medical diagnosis (e.g., getting cancer), then presents the population base rates of that diagnosis, and then asks participants their new likelihood (*41*). Thus, it is unclear whether optimistic learning biases (*15*, *26*, *28*–*30*) emerge in contexts where outcomes are both self-relevant and directly experienced. Last, while an emerging literature links variation in learning rates to internalizing disorders (*15*, *57*, *58*), prior studies have not assessed whether premorbid risk factors for internalizing psychopathology, such as NE, predict individual differences in learning. Thus, it remains an open question whether PE learning differences are simply a phenotype of an extant internalizing disorder or an indicator of risk.

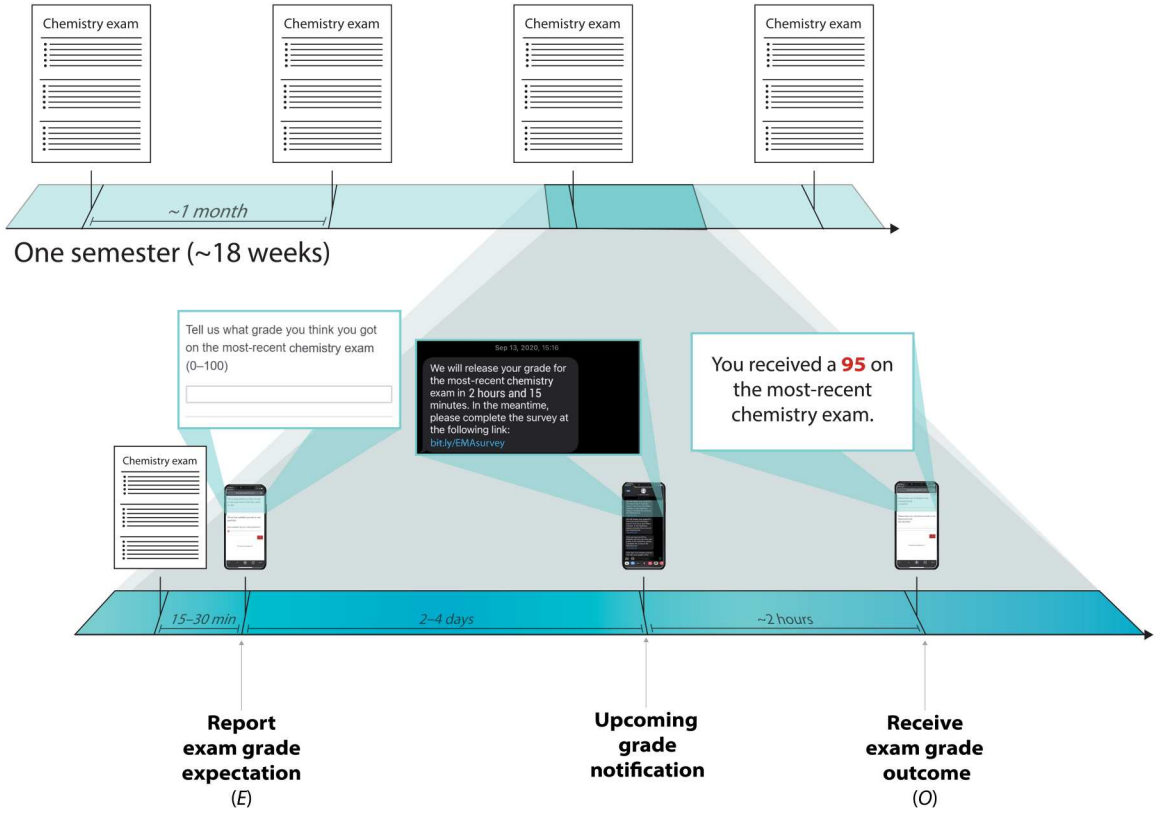

Fig. 1. Mobile phone sampling of exam grade expectations. Over the course of a single academic semester, participants in multiple cohorts reported their expected grades on four or five major chemistry exams immediately after taking exams but before viewing their grades. Grade expectations (*E*) were subtracted from grade outcomes (*O*) to compute grade PEs, which were hypothesized to drive changes in grade expectations between exams.

Accordingly, the present study investigates how personally relevant PEs and individual differences in NE affect naturalistic PE learning. We assessed in a sample of 625 undergraduate students whether midterm exam grade PEs drove updates to future grade expectations. Using a cell phone–based ecological momentary assessment (EMA) paradigm from a prior study measuring emotional responses to exam grade PEs (*48*), we sampled participants' expected grades following each exam but before grades were released. We operationalized PEs as the difference between expected and actual grades and hypothesized that grade PEs would engender learning, causing students to update their expectations for future exams (Fig. 1) in line with PEs. We define PE-driven expectation updating as the process by which students learn to more accurately predict future grades and the accuracy of future expectations as the outcome of this learning process. In contrast to prior RL studies using low-stakes financial outcomes, exam grade expectations are personally relevant and thus may be more subject to biased updating and motivated reasoning (*34*, *37*). However, relative to the personally relevant but ambiguous feedback used in most belief updating studies, exam grade PEs are also unambiguous learning signals and thus may be more likely to drive rational, unbiased updating (*34*). This setting allows us to specifically test whether individuals update more optimistically or more rationally following unambiguous but personally relevant outcomes.

Here, we demonstrate that students learned to predict their grades more accurately over just four exams by updating their expectations for future exams in accordance with prior grade PEs. Consistent with models of motivated reasoning (*34*–*37*, *39*), students updated their expectations optimistically on average, making larger updates after positive PEs relative to negative PEs. Moreover, individuals with elevated NE were less accurate in their exam grade expectations. Critically, this difference in accuracy was attributable to both an elevated sensitivity to positive and negative PEs, and a defensively pessimistic tendency to make negative updates to small positive PEs among people with high NE. Critically, using longitudinal measures of anxiety symptoms, we found that inaccurate expectations resulting from these differences in learning predicted the future development of anxiety symptoms in individuals with elevated NE. Our results suggest that a sensitivity to unexpected outcomes and a preference for avoiding negative PEs may lead someone to an inaccurate and pessimistic model of the world, perhaps increasing future risk for anxiety.

RESULTS

Learning from real-world PEs

Expectations become more accurate over time

Assuming that participants learned from their grade PEs, we hypothesized that the accuracy of their grade expectations would improve over time. To test whether this was the case, we operationalized expectation accuracy as the inverse of participants' unsigned PEs [i.e., 100 - absolute value of PEs; the magnitude of "surprise" (*14*)] for each exam and estimated its linear trend over exams. Participants' expectations became more accurate with each exam [B_{exam} = 1.11 (0.14), $P < 0.0001$; Fig. 2A].

Expectation updating scales with PE magnitude

To determine whether participants used PEs to improve expectation accuracy, we tested whether changes to expectations between exams were predicted by grade PEs. Participants changed their grade expectations as a function of their preceding grade PEs [B = 0.56 (0.02), $P < 0.0001$; Fig. 2B] such that higher positive PEs yielded larger increases in expectations and vice versa. We used model comparison to determine the combination of indicators providing the best fit to changes in expectation. Models included varying combinations of predictors that we hypothesized could influence expectation changes, including (i) exam grades, (ii) between-exam changes in grades ($grade_j - grade_{j-1}$; termed "change in grade"), and (iii) grade PEs. As indicated in Table 1, the model including grade PE and change in grade performed best. While the influence of grade PEs on expectations remained significant even when controlling for changes in grades, the effect of changes in grades on expectation updating was also significant [B = 0.62 (0.02), $P < 0.0001$]. Thus, PE-driven updating was a key mechanism that enabled individuals to become more accurate in their expectations over the four trials.

Updating rates are larger following positive relative to negative PEs

In the RL and belief updating literatures, findings are heterogeneous as to whether individuals learn preferentially from positive or negative PEs (*15*, *24*, *26*–*30*, *32*, *98*). Prior attempts to resolve this question have yielded contradictory results, with some studies finding support for optimistic learning biases (*15*, *26*, *28*–*30*, *98*) and a smaller subset of studies finding pessimistic biases (*24*, *27*, *32*). Here, to determine whether participants were more sensitive to PEs of a particular valence, we decomposed grade PEs into two terms: a two-level factor representing PE sign (positive versus negative) and a continuous variable representing PE magnitude (absolute value of PE). We then tested whether these terms interacted significantly, which would indicate greater sensitivity to PEs

Table 1. Model comparison of variables predicting change in grade expectation. Models that used grade PE and change in grade (within participant and between exams) yielded the optimal fit to the data.

Model	Predictors	Number of Parameters	Akaike Information Criterion	Bayesian Information Criterion
1	Grade PE	5	14,595	14,622
2	Grade	5	14,686	14,713
3	Change in grade	5	14,294	14,321
4	Grade + grade PE	6	14,195	14,228
5	Grade + change in grade	6	14,287	14,320
6	PE + change in grade	6	13,650	13,682

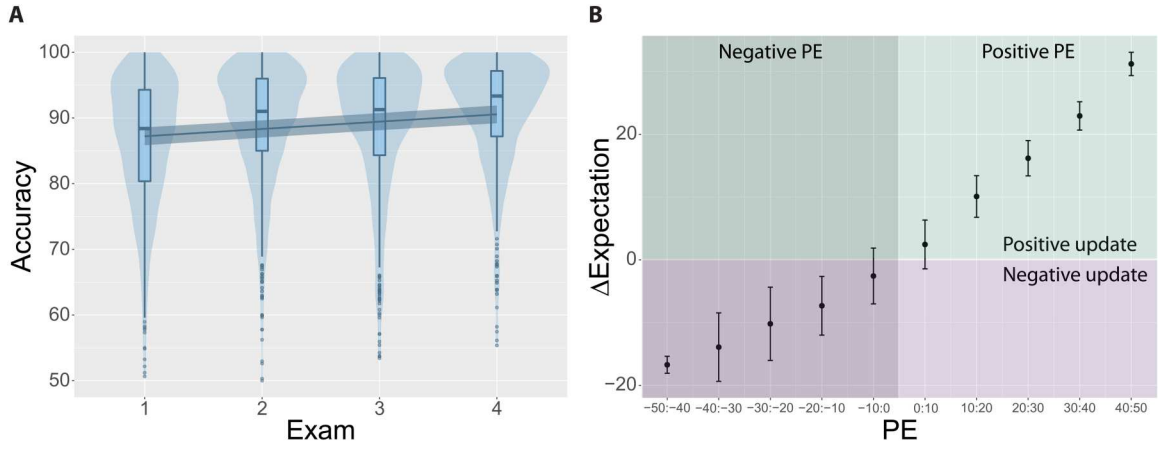

Fig. 2. PE learning drives improvements in expectation accuracy. (**A**) Over time, the accuracy of participants' grade expectations improved. This suggests that over just four exams, participants learned to better predict their grades. (**B**) Computing PEs as the differences between participants' exam grade expectations and their actual grades, reveals that PEs prompt updates to expectations, with positive PEs yielding larger changes relative to negative PEs. Taken with the observed trend of improving expectation accuracy over time, this constitutes evidence of PE-driven learning.

of one valence but not the other. In line with findings of optimistic learning biases, participants made larger expectation changes following positive PEs relative to negative PEs [B = 0.29 (0.04), P < 0.0001; Fig. 2B]. Every one-point increase to a positive PE led to increases in grade expectations of 0.80 points. In contrast, every one-point decrease to a negative PE led to decreases in grade expectations of only 0.38 points. These results indicate that individuals learned preferentially from positive PEs, mirroring findings from some laboratory-based studies of PE learning (*15*, *26*, *29*).

Individual differences in NE are linked to variability in real-world learning

NE is associated with pessimistic and inaccurate expectations regardless of outcome

Prior work suggests that differences in PE learning may affect the ability of anxious (*57*, *58*) and depressed individuals (*15*, *16*, *59*, *60*) to maintain accurate expectations in the face of unexpected, negative outcomes. For individuals with internalizing disorders, inaccurate expectations for the future may promote distress and impairment in daily life (*18*). However, it may be the case that differences in PE learning are not an emergent symptom of internalizing disorders but rather a risk factor for their development (*83*, *84*). Here, we evaluated whether NE, a personality trait that predicts the future development of anxiety (*75*, *80*, *99*), might increase psychiatric risk via differences in PE learning. Despite NE not predicting differences in actual exam grades [effect of NE on grade: B_{NE} = −0.092 (0.093), P = 0.33], individuals with higher NE were systematically more pessimistic in their expectations [effect of NE on grade expectation: B_{NE} = −0.32 (0.081), P < 0.0001]. As a result, relative to those with low NE, high-NE individuals were consistently less accurate when predicting their exam grades [effect of NE on expectation accuracy: B_{NE} = −0.089 (0.039), P = 0.022; Fig. 3B and fig. S1] and experienced more positive PEs [effect of NE on PE: B_{NE} = 0.236 (0.065), P = 0.0003; Fig. 3A and fig. S1], whereas PEs varied more randomly in valence (positive versus negative PEs) for lower-NE individuals across exams.

NE is linked to more reactive expectation updating

Given emerging evidence that internalizing psychopathology is linked to variation in learning rates (*15*, *18*, *20*, *57*, *58*), we hypothesized that the impaired expectation accuracy in high-NE individuals might result from differences in expectation updating. Compared to participants with lower NE, participants with high levels of NE made larger changes to their expectations after equivalently sized PEs [$B_{NE \times PE}$ = 0.010 (0.003), P = 0.019; Fig. 3D], evidencing an overall more reactive updating style. Broadly, this led high-NE individuals to overcorrect their expectations following PEs, which is one mechanism by which high-NE individuals were less accurate in their expectations. This finding mirrors prior work, suggesting that exceedingly high PE learning rates can prevent an individual from achieving accurate expectations in the face of small (i.e., noisy) learning signals (*23*, *63*).

Individuals higher in NE do not display differences in expectation updating to positive versus negative PEs

While at the level of the entire sample we observed an optimistic bias in which individuals asymmetrically updated their expectations following positive versus negative PEs, some work suggests that these biases are reversed in those with internalizing disorders (*16*, *59*, *60*). However, low- and high-NE individuals did not exhibit different valence-based updating asymmetries, as indicated by a lack of a significant linear interaction between NE, PE sign, and PE magnitude [$B_{NE \times PE\ sign \times PE\ magnitude}$ = 0.01 (0.01), P = 0.45].

While the preceding results revealed no differences in the way higher- versus lower-NE individuals learned from positive versus negative PEs, it is also possible that the linear functional form for the interaction limited our ability to detect differences in updating as a function of PE valence. Therefore, we further explored whether NE predicted valence-dependent asymmetries in expectation updating by specifying a multilevel Bayesian model with b-splines, which permitted nonlinearity in expectation updating following positive versus negative PEs. Results from this nonlinear model are detailed in the Supplementary Materials and provide some evidence that the optimistic updating bias observed at the group level may be attenuated in individuals with elevated NE (see the

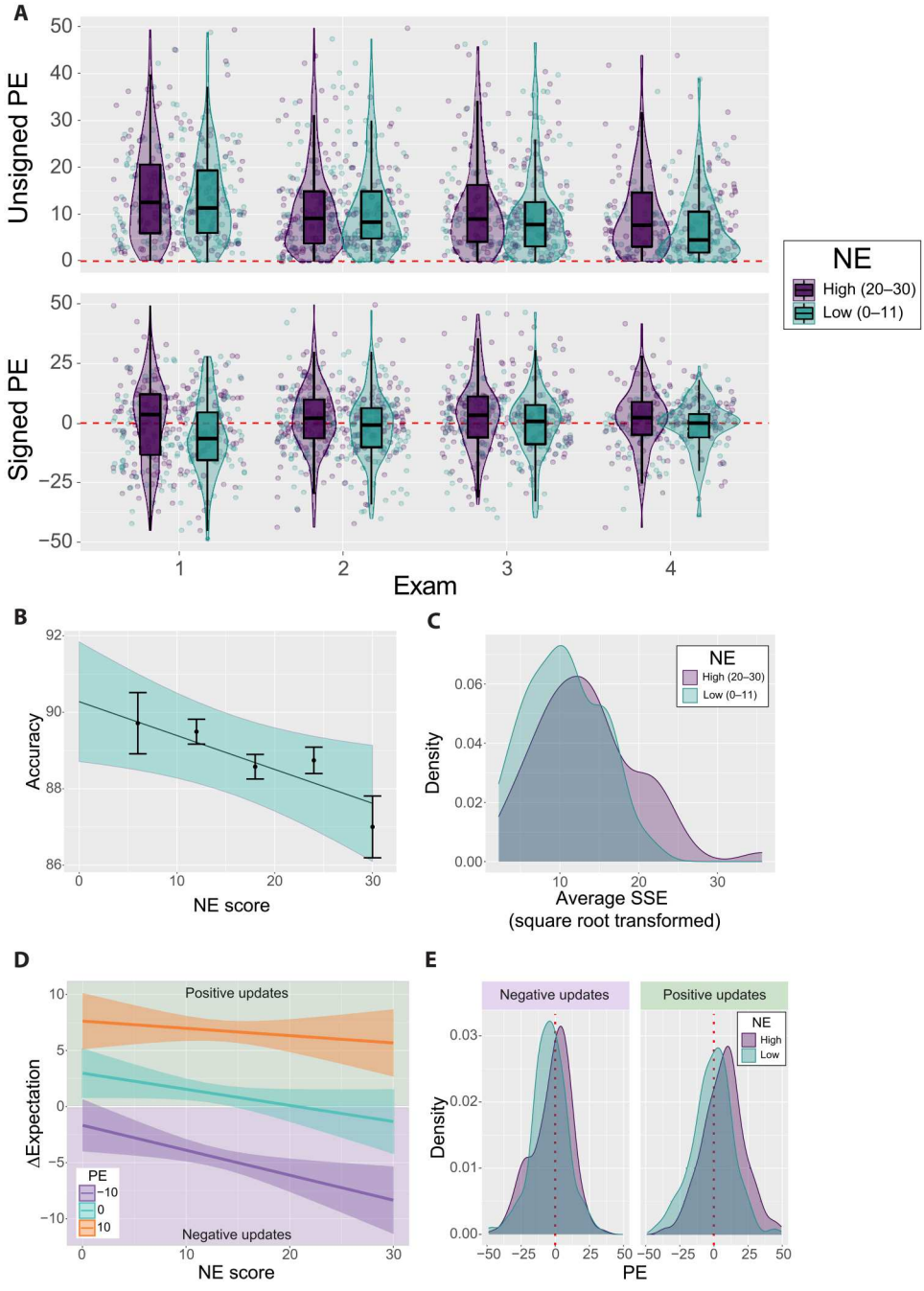

Fig. 3. NE is linked to poor accuracy, hyperreactive updating, and defensive pessimism. (**A**) Individuals with elevated NE were more pessimistic, which led to more positive PEs and larger unsigned PEs—that is, lower accuracy—over time. (**B**) Individuals with elevated NE reported less accurate expectations than their lower-NE counterparts. (**C**) Lower accuracy in higher-NE individuals is further reflected in greater average sum of squared error (SSE) scores relative to lower-NE individuals. (**D**) Individuals with elevated NE not only made larger changes to expectations following PEs but also tended to make pessimistic updates even when PEs were equal to zero, consistent with defensive pessimism. (**E**) Another signature of defensive pessimism was that higher-NE individuals tended to require larger positive PEs to update expectations positvely.

"Nonlinear valence asymmetry model" section in the Supplementary Text; fig. S2).

NE is linked to reduced accuracy when reducing expectations

Having demonstrated that high-NE individuals had less accurate expectations, we next tested whether the more reactive updating style of higher-NE individuals accounted for such accuracy deficits. Accuracy varied as a function of expectation updating across levels of NE [$B_{NE \times update}$ = 0.01 (0.002), P = 0.00014; fig. S3] such that individuals with elevated NE were less accurate after making larger updates. However, the effects of updating on accuracy differed depending on whether individuals were increasing or decreasing their expectations. Specifically, higher-NE individuals were

significantly less accurate when reducing their expectations, relative to low-NE individuals [$B_{NE \times update\ sign}$ = -0.26 (0.08), P = 0.0011]. Thus, while high-NE individuals displayed a more reactive updating style irrespective of PE sign, their updated expectations were less accurate specifically when they reduced expectations.

In sum, higher-NE individuals have a reactive updating style, but their overall poorer accuracy emerges when they reduce their future expectations. Moreover, unlike lower-NE individuals who tend to increase expectations following positive PEs and reduce expectations following negative PEs, higher-NE individuals tend to make negative updates after accurate expectations (PE = 0; Fig. 3D) and even after small positive PEs. Such a defensively pessimistic tendency is evident in the distribution of PEs that yielded positive and negative updates in high- versus low-NE individuals (Fig. 3E). Whereas low-NE individuals required larger negative PEs than high-NE individuals to justify a negative update, consistent with the optimism bias observed at the sample level, higher-NE individuals required relatively small negative PEs to decrease their expectations. To decrease their expectation on the next exam, the statistical model indicated that lower-NE individuals (i.e., bottom NE quartile) required a negative PE larger than −2.52, while higher-NE individuals (top NE quartile) only required a negative PE of −0.16 to lower their expectations. Thus, a higher-NE individual's impaired accuracy following negative updates may be the result of both overly large updates and a greater tendency to make negative updates that were incongruent with PE sign.

NE is linked to reduced accuracy following small PEs

The foregoing results suggest that individuals high in NE were more sensitive to PEs during learning and tended to make defensively pessimistic updates even when such updates may not have been explicitly warranted (e.g., when PEs were zero or positive). Thus, we tested whether specific PE features (e.g., PE magnitude or PE sign) were linked not just to differences in updating but to less accurate expectations in high-NE individuals. In fact, individuals with high levels of NE were less accurate after experiencing small PEs [$B_{NE \times PE\ magnitude}$ = 0.01 (0.0041), P = 0.02] but not after experiencing positive or negative PEs in particular [$B_{NE \times PE\ sign}$ = −0.01 (0.04), P = 0.71]. The predicted effects from this model are visualized in Fig. 4, where PEs greater than 10 (i.e., a difference of at least one letter grade) are categorized as "large" and PEs less than or equal to 10 are categorized as "small." Thus, accuracy impairments in higher-NE individuals were most prominent after small PEs and are likely due to differences in updating style, including both defensive pessimism and hyperreactive updating. Compared to large PEs, small PEs are more ambiguous learning signals, may be more open to interpretation, and thus may permit more biased updating, leading higher-NE individuals to maintain a less accurate and more pessimistic model of the world.

Poorer expectation accuracy mediates the link between NE and long-term anxiety symptoms

As noted in Introduction, individuals with heightened NE are at risk for internalizing disorders and, specifically, anxiety symptoms (*82*). One implication of the finding that learning differences among high-NE individuals yield less accurate expectations is that such an inaccurate and pessimistic model of the world may predict downstream anxiety risk. As a preliminary test of this hypothesis, we recontacted previously enrolled participants to measure their current anxiety symptom severity [n = 364; approximately 6 to 36 months after initial enrollment; measured using the Generalized Anxiety Disorder 7-Item Scale (GAD-7)] (*100*). We then tested whether individual differences in the overall accuracy of one's grade expectations mediated the pathway between NE and long-term anxiety symptoms, measured at follow-up. To do this, we calculated a participant's average expectation accuracy over the semester [quantified as the average sum of squared error in expectation over exams (average SSE)] as an indicator of these learning differences and tested whether average SSE mediated the links between baseline NE and the development of future anxiety symptoms. To ensure the reliability of the average SSE metric, average SSE was only calculated for participants who provided EMA data (i.e., predictions) for at least four exams. Thus, of the 364 participants who completed anxiety measures at follow-up, 191 were included in the mediation analysis.

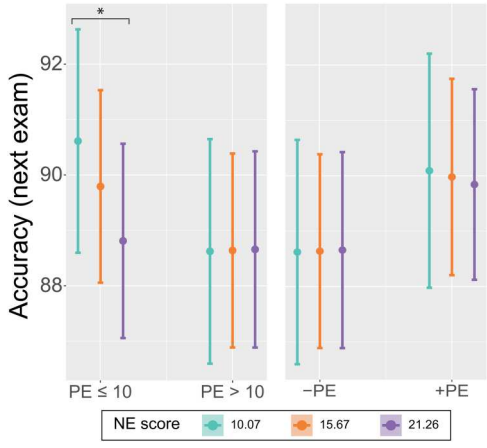

Fig. 4. NE is linked to reduced accuracy following small PEs. Comparing the accuracy of updated expectations after participants experienced large versus small and positive versus negative PEs, we found that participants with higher levels of NE made less accurate updates to their grade expectations specifically after small PEs. No significant differences were observed between PEs of opposing valence (*P < 0.05).

Before evaluating the full mediation model (Fig. 5A), we separately examined the effects from the two constituent models that comprised the full mediation (see Materials and Methods for additional details). This allowed us to understand how baseline NE, average SSE, and anxiety symptoms at follow-up were related, without yet accounting for the indirect (i.e., mediation) effect of average SSE between baseline NE and follow-up anxiety symptoms. Because of their observed differences in learning, the first model revealed that individuals with elevated NE exhibited larger average SSE values, indicative of impaired expectation accuracy [B_{NE} = 5.74 (2.13), P = 0.0076; Fig. 3C]. In the second model, anxiety scores at follow-up were significantly associated with baseline anxiety scores [$B_{Sx\ (baseline)}$ = 0.244 (0.093), P = 0.0098] and average SSE [$B_{Avg\ SSE}$ = 0.0046 (0.0018), P = 0.012]. Critically, the association between baseline NE and follow-up anxiety symptoms was also significant in this model [B_{NE} = 0.199 (0.070), P = 0.0053].

Next, we evaluated these effects in the full mediation model (paths labeled in Fig. 5A). In line with the effects that we observed before including average SSE as a mediator, the full mediation model revealed a significant effect of NE on average SSE [B = 5.81; 95% confidence interval (CI): (1.68, 9.98)]. Although the effect of average SSE on follow-up anxiety symptoms did not

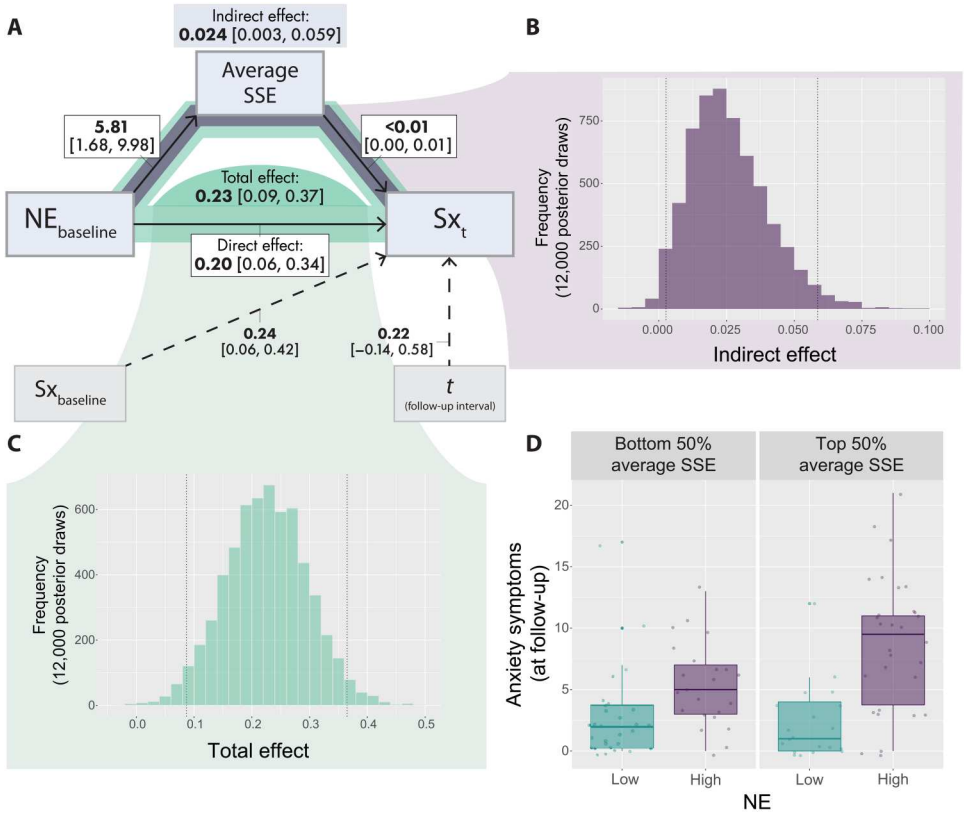

Fig. 5. Differences in learning outcomes mediate the relationship between NE and the long-term development of anxiety. (**A**) Mediation path diagram with path coefficients—average SSE mediated the relationship between NE at baseline and anxiety symptoms at follow-up (Sx_t; n = 191). (**B**) The indirect effect of NE on anxiety symptoms via average SSE was different from zero, as indicated by 95% confidence intervals for 12,000 draws from the posterior distribution. (**C**) The total effect of NE and average SSE on anxiety symptoms was also different from zero. (**D**) Larger impairments in learning outcomes, quantified as the average SSE in prediction over a given participant's exams, predicted increases in anxiety symptoms 6 to 36 months later and more so for individuals with elevated NE at baseline. Left includes participants with average SSE scores below the median, and right includes participants with average SSE scores above the median.

differ from zero in the full mediation model [$B < 0.01$; 95% CI: (0.00, 0.01)], both the direct effect [$B_{\text{direct effect}} = 0.20$; 95% CI: (0.06, 0.34)] and the total effect [$B_{\text{total effect}} = 0.23$; 95% CI: (0.09, 0.37); Fig. 5C] of NE on anxiety symptoms were significant. Moreover, average SSE significantly mediated the relationship between NE and follow-up anxiety symptoms, as indicated by a significant indirect effect [$B_{\text{indirect effect}} = 0.024$; 95% CI: (0.003, 0.059); Fig. 5B]. This suggests that while NE alone may predict future anxiety symptoms, accuracy impairments resulting from hyperreactive updating and defensive pessimism function as a potential pathway through which NE might predict such long-term increases in anxiety.

Last, to better understand how average SSE was related to anxiety symptoms at follow-up, we visualized average anxiety scores at follow-up for participants with high versus low NE (top and bottom quartiles) and average SSE in the top versus bottom 50% (i.e., median split). This visualization confirms that high-NE individuals with higher average SSE values reported the highest anxiety symptoms at follow-up (Fig. 5D). Together, this suggests that the learning differences that we observed in high-NE individuals, specifically hyperreactive updating and defensive pessimism, are not only early markers of psychopathology but also lead to impaired expectation accuracy, which, in turn, predicts future anxiety symptoms in those with elevated NE.

DISCUSSION

RL work suggests that organisms use PEs to fine-tune their expectations for the environment and accurately prepare for what is likely to come (*13*, *101*–*103*). Recent computational models formalizing the PE learning process reveal variability in the way humans learn from PEs (*24*, *27*, *32*). However, no studies to date have investigated whether such variability is present in the way humans learn from salient events in everyday life. We used a high-stakes, naturalistic situation to test whether individuals learn from unexpected and self-relevant outcomes. Unlike laboratory-based tasks that require hundreds of trials, we demonstrated in this high-stakes setting that individuals use PEs to form accurate expectations about future exam grades within just four trials. In line with some prior work (*29*, *34*, *41*), we also found evidence in support of a general optimism bias in expectation updating, whereby, in aggregate, individuals preferentially discounted negative relative to positive PEs during learning. Moreover, we found that these learning effects vary as a function of a personality profile linked to the development of anxiety disorders. These findings build on a growing literature investigating ecological human learning (*95*, *96*, *104*) and underscore the importance of using real-world data both to develop ecologically valid theories of human behavior (*105*) and to improve etiological models of psychopathology.

To achieve more accurate models of the world, individuals must update their expectations for the future in accordance with two key features of PEs: the valence (i.e., whether an outcome was better or worse than expected) and the magnitude (i.e., the degree of surprise associated with an outcome) (*1*, *12*–*14*). While research in the laboratory implicates PE valence and magnitude as drivers of expectation updating (*23*, *27*, *43*), we extend these findings to a naturalistic context by demonstrating that PE valence and magnitude are also primary drivers of expectation updating following personally meaningful events in daily life. Moreover, our findings indicate that the same PE learning mechanisms observed in highly controlled laboratory settings enabled individuals to develop more accurate expectations over just four real-world trials.

An important disparity between value-based RL and motivated belief updating literatures involves the necessity of forming accurate expectations. Value-based RL tends to assume that accurate predictions are central (*12*–*14*), whereas individuals tend to hold optimistic and often inaccurate motivated beliefs (*34*–*37*, *39*). However, while value-based RL tasks commonly use low-stakes learning signals, motivated belief tasks use personally meaningful signals, which may be more open to interpretation (*34*). Here, using unambiguous exam grade PEs, we specifically tested whether individuals update more optimistically or more rationally following unambiguous but still ego-relevant outcomes. Investigating PE-based learning in this naturalistic setting enabled us to bridge the gap between lower-order, value-based expectation updating in RL and higher-order, self-relevant belief updating. Overall, we found evidence of a general asymmetry in how much individuals updated their expectations following positive versus negative PEs. Consistent with suggestions that humans display positivity biases in belief updating (*15*, *26*, *28*–*30*, *41*), our group-level results indicated that participants discounted negative PEs and learned preferentially from positive ones. Given that exam grade expectations contain signals pertaining to self-relevant dimensions such as intelligence and ability, the emergence of optimistic updating in this setting underscores the role of motivated reasoning as a driver of biased updating (*34*, *35*, *37*, *38*). However, there are likely contexts in which learning rates are higher for positive PEs and others in which learning rates are higher for negative PEs. Negative life events may be more impactful than positive ones (*106*), and in some contexts, this is undeniable. Highly salient negative PEs, such as traumatic events, can engender one-shot learning, resulting in massive changes to higher-order beliefs and, in some cases, posttraumatic stress disorder (*97*, *107*). Thus, a simple "one-size-fits-all" account of learning rate asymmetries is unlikely to approximate the real-world PE learning process. Future learning studies in naturalistic contexts are needed to understand the contextual moderators of valence-based updating asymmetries.

While learning rates commonly vary within and between persons (*17*, *18*, *23*, *42*), individual differences in learning rates have been linked to internalizing disorders (*15*, *18*, *20*, *57*, *58*, *63*, *108*), and it has been suggested that such individual differences may contribute to psychopathology development (*18*). We found that NE, a personality trait that predicts the development of internalizing disorders (*67*, *75*, *80*, *99*), modulated real-world, high-stakes PE learning, such that individuals with elevated NE updated their expectations at greater rates than those with lower NE and made irrational negative updates congruent with defensive pessimism. Over time, our findings highlight how inaccurate expectations resulting from these learning differences may predict the development of long-term anxiety symptoms in vulnerable individuals with elevated NE. Together, these results extend and challenge extant work implicating differences in learning rates in fully developed psychiatric disorders (*10*, *58*) and suggest differences in learning that may emerge from maladaptive beliefs and ultimately predict the development of anxiety in higher-NE individuals.

Clinical theories linking RL to psychopathology emphasize individual differences in PE learning as a mechanism that yields inaccurate expectations for the future, which, in turn, may promote distress and impairment in daily life (*18*). Although individuals with elevated NE in our sample did not score differently on exams, they reported significantly lower expectations on average, consistent with defensive pessimism (*52*, *109*–*111*), and, critically, were less accurate predictors of their exam grade performance. Thus, we hypothesized that impaired expectation accuracy in higher-NE individuals resulted from differences in expectation updating rates relative to their lower-NE counterparts.

While individuals with elevated NE tended to make larger updates to expectations, indicative of a general sensitivity to PEs, they did not exhibit different valence-based updating asymmetries—learning differences commonly observed in depressed (*53*, *54*) and anxious individuals (*10*, *20*, *58*). Nonetheless, individuals with elevated NE were less accurate after making negative updates to their expectations, which was due to a defensively pessimistic tendency to make negative updates that did not respect the valence of prior PEs. Critically, for high-NE individuals, we found that accuracy deficits were most pronounced following small PEs. While further research is necessary, this may suggest that this defensively pessimistic bias is more likely to occur when PEs are small and thus represent more ambiguous learning signals, as some prior work suggests (*34*). The fact that higher-NE individuals did not exhibit differences in accuracy after experiencing large PEs may support this idea. Following large PEs, participants paid similar respect to the signal strength of PE learning signals, perhaps because larger PEs are more difficult to ignore (*34*), and failure to learn from them might portend similarly large PEs and greater uncertainty in the future.

While additional follow-up work is warranted, defensive pessimism, combined with the tendency to overlearn from PEs, may result from an increased sensitivity to events that are perceived as unpredictable—a cognitive trait that predominates in clinical models of NE (*82*, *112*). It may be that individuals with elevated NE possess a greater sensitivity to PEs and thus a lower threshold for tolerating unexpected outcomes, regardless of their valence (*113*). Moreover, given a learning history rich in negative PEs, defensive pessimism may be a conditioned adaptation aimed at mitigating the impact of such events in the future. Thus, the learning differences we observed in individuals with elevated NE support the notion that a greater sensitivity to unpredictability might confer sensitivity to PEs and bias the way information is weighted during learning. More broadly, the present findings suggest that risk for anxiety might emerge from how higher-NE individuals update their expectations, ultimately leading to pessimistic, inaccurate models of the world.

However, this study is not without limitations. Because the order of midterm exams was not randomized between participants, we cannot rule out the possibility that improving accuracy over time was an outcome of latter exam grades being easier to predict and

not just PE-based learning. Future investigations should rule out such a possibility by replicating our analyses across a range of academic classes with greater variability in the content and timing of exams. Furthermore, this type of naturalistic learning study should be replicated in a nonstudent sample to ensure that the effects are not limited to this single unique context.

In conclusion, we provide evidence of PE-based learning in a real-world, high-stakes context that can be observed over just a handful of trials. In aggregate, individuals displayed greater updating after positive relative to negative PEs, consistent with the role of motivated reasoning in optimistic updating. However, individuals with a personality phenotype linked to anxiety disorders displayed key differences in learning, which caused them to be more inaccurate in their expectations and, in turn, predicted long-term increases in anxiety symptoms. Given such a diathesis (*114*), we hypothesize that a conditioned aversion to negative and unpredictable events would lead a person to develop a pessimistic and inaccurate model of the world, which may predict risk for anxiety.

MATERIALS AND METHODS

Participants

Participants were 740 undergraduate students recruited from chemistry classes at the University of Miami between August 2019 and December 2020. Prior to enrollment, participants provided informed consent per study protocol approved by the Institutional Review Board at the University of Miami (IRB# 20180529). Over three semesters (fall 2019, n = 187; spring 2020, n = 315; fall 2020, n = 436; with 198 students enrolled in more than one semester), students in three different chemistry courses (general chemistry, organic chemistry 1, and organic chemistry 2) participated in a semester-long EMA study that assessed exam grade expectations for the four to five midterm exams in each class (Fig. 1). Given that exams occurred in a real-world university class, the order of exams was not randomized, but exams were free to vary in content between cohorts.

Exclusion criteria

Participants who did not participate sufficiently in EMA sampling (i.e., provide grade predictions for at least two consecutive exams) were excluded from the final analysis sample (n = 115 participants excluded). This yielded a final analysis sample of 625 participants. Demographic characteristics for the full sample and the final analysis sample are presented in Table 2.

Table 2. Demographic characteristics of full sample and analysis sample. Note that participants who did not complete EMA surveys (n = 115) following at least two exams were excluded from the final analysis sample.

Demographic	Full sample	Analysis sample
n	**740**	**625**
Gender (%)		
Female	74.19	75.36
Male	25.81	24.64
Race (%)		
White or Caucasian	66.62	69.28
Black or African American	10.27	9.12
Asian or Asian American	12.43	12.16
Native American	0.41	0.16
Native Hawaiian or Pacific Islander	0.14	0.16
Multiracial	7.70	7.20
Other	2.43	2.08
Ethnicity (%)		
Non-Hispanic	71.62	71.52
Hispanic/Latino	28.38	28.48

Experimental design

Initial laboratory sessions

At the start of academic semesters, students interested in the study participated in an initial laboratory session during which they provided informed consent and authorized the study team to access their exam grades from course professors. Specifically, before exam grades were posted for students to view, chemistry professors provided the study team with the exam grades of study participants. Participants provided contact information for the distribution of EMA surveys and were informed that surveys would be distributed to their mobile phones via text messages [short message service (SMS)] containing Uniform Resource Locator (URL) links to the Qualtrics online survey platform (*115*). Thus, all participants were required to have a cell phone capable of internet access and receiving text messages. To incentivize the completion of EMA surveys, participants were compensated with course extra credit proportional to their EMA completion rates.

Baseline questionnaire battery. During the initial laboratory visit, participants completed a baseline questionnaire battery that assessed demographics, personality traits, and psychopathology symptoms, including generalized anxiety. Individual differences in NE were derived from participants' scores on the Big-Five Inventory, Extra Short Version (BFI) (*116*), and baseline anxiety symptoms were derived from the participants′ scores on the GAD-7 (*100*).

Assessment of exam grade expectations

Within 15 to 30 min of the end of each midterm exam period, researchers sent a SMS to participants requesting that they report the grade they expected to receive on that exam. Grade expectations were entered into a survey text box. Only numeric responses between 0 and 100 were accepted, and participants were automatically prompted to reenter their expected grade if their response was not within this range.

Release of midterm exam grade outcomes

After receiving exam grades from chemistry professors, participants were notified via SMS that their grades were ready to be viewed before being posted on the course website. To view exam grades, participants clicked on the URL that they received, which led them to a webpage requesting that they enter their contact information (i.e., last name and cell phone number). This information was automatically cross-referenced with our participant database, and participants were redirected to a webpage containing their exam grade (e.g., "You received a 65 on the most-recent chemistry exam").

Measurement of longitudinal anxiety symptoms

Anywhere from 6 to 36 months after participating in this study, participants who consented to be contacted in the future were offered a

financial incentive (Amazon gift card) to complete a follow-up measurement of their current anxiety symptoms. Similar to the baseline questionnaire battery, follow-up anxiety scores were measured using the GAD-7.

Preprocessing and calculation of learning variables

Outcome and prediction-related variables

Exam grade PEs. Exam grade PEs were computed as the difference between participants' expected grades for an exam and the actual grades they received on that exam

$$\mathrm{PE}_{ij} = O_{ij} - E_{ij} \quad (1)$$

where i denotes observations for a given participant, j denotes observations for one of the exams, E represents an exam grade expectation, and O represents the exam grade outcome that one received.

Indices of learning

Expectation accuracy. We hypothesized that learning would manifest as an increase in the accuracy of exam grade predictions over the four to five exams. To determine whether participants learned over the semester, we operationalized the outcome of learning as the expectation accuracy, using the unsigned PE [i.e., magnitude of surprise (*14*)]. We subtracted unsigned PEs from 100 to invert the sign such that greater accuracy corresponded to higher values of the variable

$$\mathrm{Accuracy}_{ij} = 100 - \mid O_{ij} - E_{ij} \mid = 100 - \mid \mathrm{PE}_{ij} \mid \quad (2)$$

Expectation updates (between exams). Improvements in expectation accuracy, the outcome of learning, require changes in one's expectations, which constitute the process of learning. We computed between-exam changes in expectation, termed expectation updates (ΔE), as the difference between grade expectations for consecutive exams

$$\Delta E_{ij} = E_{ij} - E_{ij-1} \quad (3)$$

Average error in prediction (average SSE). To quantify individual participants' relative ability to accurately predict their exam grades over time, we computed a metric representing their average error in prediction as the sum of squared errors in prediction, normalized by the number of observations included in the calculation (i.e., the number of exams for which a participant reported a PE)

$$\text{Average SSE}_i = \left(\sum_{j=1}^{n} \mathrm{PE}_{ij}{}^2\right) * 1/n \quad (4)$$

To ensure the reliability of average SSE estimates, average SSE was only calculated for participants who reported predictions (and thus PEs) for at least four exams.

Statistical analysis

Statistical analyses were conducted using the R programming language (*117*). Distributions of all variables were assessed for normality, and descriptive statistics for each variable were extracted before statistical modeling. Given the hierarchical structure of the dataset (i.e., exams within participants and within cohorts), we used multilevel regression models to account for participant-specific and cohort-specific effects. All linear mixed-effects models were specified and evaluated using the "lme4" package in R (*118*), and Bayesian mixed-effects regression models were evaluated using the "brms" package (*119*). To ensure accuracy in our estimation of learning effects, PEs and expectation updates greater than 50 or less than −50 were censored before testing learning models. This resulted in 30 observations being excluded from our dataset. Censoring these outlying observations from our analyses did not alter the learning effects described in the following (i.e., in Eqs. 5 and 6). Moreover, logistic regression results suggest that this censoring was unrelated to participants' NE scores [$B_{\mathrm{NE}} = 0.96$ (0.154), $P = 0.533$].

Learning from exam grade PEs

Do grade expectations become more accurate over time? To test whether participants' grade expectations became more accurate (constituting evidence of learning), we specified a linear mixed-effects model in which expectation accuracy was regressed onto trial number (j)

$$\mathrm{Accuracy}_{ij} \sim j + (1 \mid \mathrm{cohort}/i) \quad (5)$$

where j represents exam, i defines individual participants as random-effects levels, and the term "cohort" accounts for the dependencies between participants who participated within the same semester (i.e., cohort). We hypothesized that over the course of the semester, as experience taking exams accrues, participants become increasingly accurate in their exam grade expectations.

Do PEs drive updates to expectations? In RL frameworks, PEs function as learning signals to update expectations in order to derive a more accurate model of the world (*12*, *13*, *101*, *120*). We thus tested whether a PE on one exam predicted changes in exam grade expectations on the next exam

$$\Delta E_{ij} \sim \mathrm{PE}_{ij-1} + \Delta O_{ij} + (1 \mid \mathrm{cohort}/i) \quad (6)$$

We hypothesized that exam grade PEs would be positively associated with updates to exam grade expectations, which would support the hypothesis that PEs function as learning signals. We consider the parameter estimate for the $\mathrm{PE}_{ij\ -\ 1}$ term to be similar to a learning rate (*12*), representing the magnitude of an expectation update given a PE.

Because individual students' exam grades fluctuate over the semester, expectation updates will vary not only because of learning but also because of changes in the difficulty ofcourse material or course engagement. To account for this, in all models, we confirmed that PE learning is present when accounting for within-participant changes in grades between exams ($\Delta O_j = O_{j\ +\ 1} - O_j$). As a follow-up test, we conducted a model comparison to rule out alternative explanations for updates to exam grade expectations. Competing models included combinations and subsets of the following set of linear regressors: grade PE, grade expectation, and change in grade (ΔO). This model comparison indicated that a model containing both grade PE and change in grade predictors yielded the best fit to the data and was thus the foundation for subsequent models predicting expectation updating.

Do updating rates differ by PE valence? Because existing work indicates that learning rates differ between positive and negative PEs, we specified an additional multilevel model in which separate terms representing PE valence (categorical variable) and PE magnitude (continuous variable; i.e., unsigned PE) were added to a model

predicting expectation updates

$$\Delta E_{ij} \sim \text{PE valence}_{ij-1} * \text{PE magnitude}_{ij-1} + \Delta O_{ij} + (1 \mid \text{cohort}/i) \quad (7)$$

Individual differences in NE

Does NE predict variability in learning outcomes? We initially tested whether NE predicted differences in the outcome of learning: expectation accuracy. To address this question, we specified a univariate mixed-effects model in which expectation accuracy was regressed onto participant-specific NE scores

$$\text{Accuracy}_{ij} \sim \text{NE}_i + j + (1 \mid \text{cohort}/i) \quad (8)$$

where NE_i represents individual participants' NE scores measured using the BFI at baseline. The exam index j was included as a covariate to account for trends in expectation accuracy over time.

Does NE predict variability in PE-based learning (i.e., updating rates)? To test whether individuals with elevated NE demonstrated systematic differences in PE-driven learning, we first tested whether NE moderated the impact of PE on changes to expectation. As above, we included change in grade as a covariate

$$\Delta E_{ij} \sim \text{PE}_{ij-1} * \text{NE}_i + \Delta O_{ij} + (1 + \text{NE}_i \mid \text{cohort}/i) \quad (9)$$

This model included a random slope for NE to account for between-participant differences in the relative impact of NE on expectation updating.

Does NE generate differences in learning outcomes by way of differences in expectation updating? To test whether the process of PE-driven learning resulted in less accurate expectations for individuals with elevated NE, we tested a linear mixed-effects model in which expectation accuracy at the next exam was predicted by the interaction between NE and expectation updates. As in the preceding models, we controlled for trends in expectation accuracy over time

$$\text{Accuracy}_{ij+1} \sim \text{NE}_i * \Delta E_{ij} + j + (1 \mid \text{cohort}/i) \quad (10)$$

Does NE predict valence-dependent asymmetries in expectation updating? We then tested whether changes in expectation as a function of NE differed by PE valence (i.e., positive versus negative PEs). We specified a linear mixed-effects model in which changes to expectations were predicted by a three-way linear interaction between PE magnitude, PE valence, and NE

$$\Delta E_{ij} \sim \text{PE magnitude}_{ij-1} * \text{PE valence}_{ij-1} * \text{NE}_i + \Delta O_{ij} + (1 + \text{NE}_i \mid \text{cohort}/i) \quad (11)$$

Do individuals with elevated NE learn differently from PEs of varying magnitude and valence? We lastly tested what might be driving differences in expectation accuracy among high-NE individuals. It may be that individuals with higher levels of NE became less accurate due to small or large PEs or positive versus negative PEs. To evalutethese possibilities, we regressed expectation accuracy at exam $j + 1$ onto terms representing the PE magnitude and valence at exam j. In separate two-way interactions, NE was specified as a moderator of PE valence and magnitude terms to determine whether individual differences in NE predicted variability in accuracy following PEs of differing magnitude and valence

$$\text{Accuracy}_{ij+1} \sim \text{PE magnitude}_{ij} * \text{NE}_i + \text{PE valence}_{ij} * \text{NE}_i + j + (1 + \text{NE}_i \mid \text{cohort}/i) \quad (12)$$

This model enabled us to determine whether the effects of PE valence and magnitude on the accuracy of updated expectations varied as a function of individual differences in NE.

To visualize the predicted effects from the model depicted in Eq. 11, a categorical variant of this model was formulated in which PEs were decomposed into dichotomous groups representing small (i.e., less than 10 points) versus large PEs (i.e., greater than or equal to 10 points) and positive versus negative PEs. As in Eq. 11, we specified separate two-way interactions between NE scores and the dichotomous PE magnitude and PE valence variables, respectively.

Do differences in learning outcomes mediate the relationship between NE and the long-term development of anxiety symptoms? Last, we tested whether differences in learning and, notably, resultant impairments in learning outcomes might act as a mechanism that drives the development of anxiety symptoms in individuals with elevated NE. To test this theoretical model, we formulated and fit a Bayesian regression in which participants' average SSE mediated the causal path from NE to anxiety symptoms at a latter follow-up time point (anywhere from 6 to 36 months after the study period). Participants (n = 364) provided follow-up anxiety symptom scores, of which 191 participants had sufficient data to be included in the analyses (i.e., predictions for four or more exams). While some participants provided data at multiple follow-up time points, we strictly analyzed the change in symptoms from a participant's baseline to their last (i.e., most recent) response. Anxiety symptoms assessed at the beginning of the study period (i.e., at baseline) were included as a control variable, as was the relative duration between the end of the study and the time point when follow-up symptom measures were collected. This mediation model was built from two constituent submodels

$$\text{Sx}_{it} \sim \text{Average SSE}_i + \text{Sx}_{i\text{baseline}} + \text{NE}_i + t \quad (13)$$

$$\text{Average SSE}_i \sim \text{NE}_i \quad (14)$$

To understand the effects of NE and average SSE on follow-up anxiety symptom scores, before accounting for the potential mediating effect of average SSE, we tested the models in Eqs. 13 and 14 in separate linear regressions. To estimate the mediation effect of average SSE, these models were then jointly evaluated in a Bayesian regression. To determine whether effects in the Bayesian mediation model were significant, we took 12,000 draws from the posterior distributions and computed 95% confidence intervals for each parameter.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S4
Tables S1 to S4

REFERENCES AND NOTES

1. I. P. Pavlov, *Lectures on Conditioned Reflexes: Twenty-Five Years of Objective Study of the Higher Nervous Activity (Behaviour) of Animals* (Liverwright Publishing Corporation, 1928).
2. B. F. Skinner, *The Behavior of Organisms: An Experimental Analysis* (Appleton-Century, 1938).
3. K. Friston, Hierarchical models in the brain. *PLOS Comput. Biol.* **4**, e1000211 (2008).
4. B. Atal, Predictive coding of speech at low bit rates. *IEEE Trans. Commun.* **30**, 600–614 (1982).
5. H. Feldman, K. J. Friston, Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* **4**, 215 (2010).
6. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
7. K. Friston, Prediction, perception and agency. *Int. J. Psychophysiol.* **83**, 248–252 (2012).
8. S. Maren, M. S. Fanselow, The amygdala and fear conditioning: Has the nut been cracked? *Neuron* **16**, 237–240 (1996).
9. K. S. LaBar, J. E. LeDoux, Partial disruption of fear conditioning in rats with unilateral amygdala damage: Correspondence with unilateral temporal lobectomy in humans. *Behav. Neurosci.* **110**, 991–997 (1996).
10. S. J. Bishop, C. Gagne, Anxiety, depression, and decision making: A computational perspective. *Annu. Rev. Neurosci.* **41**, 371–388 (2018).
11. A. G. Fischer, M. Ullsperger, When is the time for a change? Decomposing dynamic learning rates. *Neuron* **84**, 662–664 (2014).
12. R. Rescorla, A. Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, in *Classical Conditioning II: Current Research and Theory* (Appleton-Century-Crofts, 1972), vol. 2, pp. 64–99.
13. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 1998), *Adaptive computation and machine learning*.
14. J. M. Pearce, G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
15. H. W. Chase, M. J. Frank, A. Michael, E. T. Bullmore, B. J. Sahakian, T. W. Robbins, Approach and avoidance learning in patients with major depression and healthy controls: Relation to anhedonia. *Psychol. Med.* **40**, 433–440 (2010).
16. D. R. Strunk, H. Lopez, R. J. DeRubeis, Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behav. Res. Ther.* **44**, 861–882 (2006).
17. J. T. McGuire, M. R. Nassar, J. I. Gold, J. W. Kable, Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* **84**, 870–881 (2014).
18. C. Gagne, O. Zika, P. Dayan, S. J. Bishop, Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife* **9**, e61387 (2020).
19. D. Meder, K. H. Madsen, O. Hulme, H. R. Siebner, Chasing probabilities—Signaling negative and positive prediction errors across domains. *Neuroimage* **134**, 180–191 (2016).
20. M. Browning, T. E. Behrens, G. Jocham, J. X. O'Reilly, S. J. Bishop, Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat. Neurosci.* **18**, 590–596 (2015).
21. S. J. Gershman, Y. Niv, Novelty and inductive generalization in human reinforcement learning. *Top. Cogn. Sci.* **7**, 391–415 (2015).
22. A. Houillon, R. C. Lorenz, W. Boehmer, M. A. Rapp, A. Heinz, J. Gallinat, K. Obermayer, "Chapter 21 - The effect of novelty on reinforcement learning," in *Progress in Brain Research*, V. S. C. Pammi, N. Srinivasan, Eds. (Elsevier, 2013); www.sciencedirect.com/science/article/pii/B9780444626042000216, vol. 202 of *Decision Making*, pp. 415–439.
23. M. R. Nassar, K. M. Rumsey, R. C. Wilson, K. Parikh, B. Heasly, J. I. Gold, Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.* **15**, 1040–1046 (2012).
24. Y. Niv, J. A. Edlund, P. Dayan, J. P. O'Doherty, Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32**, 551–562 (2012).
25. S. Palminteri, M. Lebreton, The computational roots of positivity and confirmation biases in reinforcement learning. *Trends Cogn. Sci.* **26**, 607–621 (2022).
26. M. J. Frank, A. A. Moustafa, H. M. Haughey, T. Curran, K. E. Hutchison, Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16311–16316 (2007).
27. S. J. Gershman, Do learning rates adapt to the distribution of rewards? *Psychon. Bull. Rev.* **22**, 1320–1327 (2015).
28. B. Kuzmanovic, A. Jefferson, K. Vogeley, The role of the neural reward circuitry in self-referential optimistic belief updates. *Neuroimage* **133**, 151–162 (2016).
29. G. Lefebvre, M. Lebreton, F. Meyniel, S. Bourgeois-Gironde, S. Palminteri, Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 0067 (2017).
30. C. Moutsiana, C. J. Charpentier, N. Garrett, M. X. Cohen, T. Sharot, Human frontal–subcortical circuit and asymmetric belief updating. *J. Neurosci.* **35**, 14077–14085 (2015).
31. S. Palminteri, G. Lefebvre, E. J. Kilford, S.-J. Blakemore, Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Comput. Biol.* **13**, e1005684 (2017).
32. A. Christakou, S. J. Gershman, Y. Niv, A. Simmons, M. Brammer, K. Rubia, Neural and psychological maturation of decision-making in adolescence and young adulthood. *J. Cogn. Neurosci.* **25**, 1807–1823 (2013).
33. T. Sharot, The optimism bias. *Curr. Biol.* **21**, R941–R945 (2011).
34. T. Sharot, N. Garrett, Forming beliefs: Why valence matters. *Trends Cogn. Sci.* **20**, 25–33 (2016).
35. F. Zimmermann, The dynamics of motivated beliefs. *Am. Econ. Rev.* **110**, 337–363 (2020).
36. D. Eil, J. M. Rao, The good news-bad news effect: Asymmetric processing of objective information about yourself. *Am. Econ. J. Microecon.* **3**, 114–138 (2011).
37. R. Bénabou, J. Tirole, Self-confidence and personal motivation*. *Q. J. Econ.* **117**, 871–915 (2002).
38. R. Bénabou, J. Tirole, Mindful economics: The production, consumption, and value of beliefs. *J. Econ. Perspect.* **30**, 141–164 (2016).
39. M. M. Möbius, M. Niederle, P. Niehaus, T. S. Rosenblat, Managing self-confidence: Theory and experimental evidence. *Manage. Sci.*, 10.1287/mnsc.2021.4294 (2022).
40. S. Ertac, thesis, University of California, Los Angeles (2006).
41. T. Sharot, C. W. Korn, R. J. Dolan, How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* **14**, 1475–1479 (2011).
42. R. D. Cazé, M. A. A. van der Meer, Adaptive properties of differential learning rates for positive and negative outcomes. *Biol. Cybern.* **107**, 711–719 (2013).
43. N. D. Daw, Trial-by-trial data analysis using computational models, in *Decision Making, Affect, and Learning: Attention and Performance XXIII* (Oxford Univ. Press, 2011), vol. 23.
44. K. Friston, The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
45. M. Inzlicht, B. D. Bartholow, J. B. Hirsh, Emotional foundations of cognitive control. *Trends Cogn. Sci.* **19**, 126–132 (2015).
46. M. J. Frank, B. S. Woroch, T. Curran, Error-related negativity predicts reinforcement learning and conflict biases. *Neuron* **47**, 495–501 (2005).
47. G. Hajcak, D. Foti, Errors are aversive: defensive motivation and the error-related negativity. *Psychol. Sci.* **19**, 103–108 (2008).
48. W. J. Villano, A. R. Otto, C. Ezie, R. Gillis, A. S. Heller, Temporal dynamics of real-world emotion are more strongly linked to prediction error than outcome. *J. Exp. Psychol. Gen.* **149**, 1755–1766 (2020).
49. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proc. Natl. Acad. Sci.* **111**, 12252–12257 (2014).
50. M. A. Boksem, M. Tops, A. E. Wester, T. F. Meijman, M. M. Lorist, Error-related ERP components and individual differences in punishment and reward sensitivity. *Brain Res.* **1101**, 92–101 (2006).
51. N. Cantor, J. K. Norem, Defensive pessimism and stress and coping. *Soc. Cogn.* **7**, 92–112 (1989).
52. J. K. Norem, N. Cantor, Defensive pessimism: Harnessing anxiety as motivation. *J. Pers. Soc. Psychol.* **51**, 1208–1217 (1986).
53. P. Kumar, F. Goer, L. Murray, D. G. Dillon, M. L. Beltzer, A. L. Cohen, N. H. Brooks, D. A. Pizzagalli, Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology* **43**, 1581–1588 (2018).
54. A. Mkrtchian, J. Aylward, P. Dayan, J. P. Roiser, O. J. Robinson, Modeling avoidance in mood and anxiety disorders using reinforcement learning. *Biol. Psychiatry* **82**, 532–539 (2017).
55. V. B. Gradin, P. Kumar, G. Waiter, T. Ahearn, C. Stickle, M. Milders, I. Reid, J. Hall, J. D. Steele, Expected value and prediction error abnormalities in depression and schizophrenia. *Brain* **134**, 1751–1764 (2011).
56. P. Sterzer, R. A. Adams, P. Fletcher, C. Frith, S. M. Lawrie, L. Muckli, P. Petrovic, P. Uhlhaas, M. Voss, P. R. Corlett, The predictive coding account of psychosis. *Biol. Psychiatry* **84**, 634–643 (2018).
57. L. Koban, R. Schneider, Y. K. Ashar, J. R. Andrews-Hanna, L. Landy, D. A. Moscovitch, T. D. Wager, J. J. Arch, Social anxiety is characterized by biased learning about performance and the self. *Emotion* **17**, 1144–1155 (2017).
58. J. Aylward, V. Valton, W.-Y. Ahn, R. L. Bond, P. Dayan, J. P. Roiser, O. J. Robinson, Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nat. Hum. Behav.* **3**, 1116–1123 (2019).
59. N. Garrett, T. Sharot, P. Faulkner, C. W. Korn, J. P. Roiser, R. J. Dolan, Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Front. Hum. Neurosci.* **8**, 639 (2014).

60. C. W. Korn, T. Sharot, H. Walter, H. R. Heekeren, R. J. Dolan, Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol. Med.* **44**, 579–592 (2014).
61. H. Vandendriessche, A. Demmou, S. Bavard, J. Yadak, C. Lemogne, T. Mauras, S. Palminteri, Contextual influence of reinforcement learning performance of depression: Evidence for a negativity bias? *Psychol. Med.* , 1–11 (2022).
62. A. C. Pike, O. J. Robinson, Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. *JAMA Psychiat.* **79**, 313–322 (2022).
63. H. Huang, W. Thompson, M. P. Paulus, Computational dysfunctions in anxiety: Failure to differentiate signal from noise. *Biol. Psychiatry* **82**, 440–446 (2017).
64. R. Kotov, R. F. Krueger, D. Watson, T. M. Achenbach, R. R. Althoff, R. M. Bagby, T. A. Brown, W. T. Carpenter, A. Caspi, L. A. Clark, N. R. Eaton, M. K. Forbes, K. T. Forbush, D. Goldberg, D. Hasin, S. E. Hyman, M. Y. Ivanova, D. R. Lynam, K. Markon, J. D. Miller, T. E. Moffitt, L. C. Morey, S. N. Mullins-Sweatt, J. Ormel, C. J. Patrick, D. A. Regier, L. Rescorla, C. J. Ruggero, D. B. Samuel, M. Sellbom, L. J. Simms, A. E. Skodol, T. Slade, S. C. South, J. L. Tackett, I. D. Waldman, M. A. Waszczuk, T. A. Widiger, A. G. C. Wright, M. Zimmerman, The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
65. R. C. Kessler, M. Gruber, J. M. Hettema, I. Hwang, N. Sampson, K. A. Yonkers, Comorbid major depression and generalized anxiety disorders in the National Comorbidity Survey follow-up. *Psychol. Med.* **38**, 365–374 (2008).
66. L. A. Clark, D. Watson, Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *J. Abnorm. Psychol.* **100**, 316–336 (1991).
67. AJ Shackman, JS Weinstein, SN Hudja, CD Bloomer, MG Barstead, AS Fox, EP Lemay, Dispositional negativity in the wild: Social environment governs momentary emotional experience. *Emotion (Washington, D.C.)* **18**, 707–724 (2018).
68. A. J. Shackman, D. P. M. Tromp, M. D. Stockbridge, C. M. Kaplan, R. M. Tillman, A. S. Fox, Dispositional negativity: An integrative psychological and neurobiological perspective. *Psychol. Bull.* **142**, 1275–1314 (2016).
69. K. S. Kendler, J. Kuhn, C. A. Prescott, The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *Am. J. Psychiatry* **161**, 631–636 (2004).
70. J. L. Jenness, M. Peverill, K. M. King, B. L. Hankin, K. A. McLaughlin, Dynamic associations between stressful life events and adolescent internalizing psychopathology in a multi-wave longitudinal study. *J. Abnorm. Psychol.* **128**, 596–609 (2019).
71. D. Foti, R. Kotov, D. N. Klein, G. Hajcak, Abnormal neural sensitivity to monetary gains versus losses among adolescents at risk for depression. *J. Abnorm. Child Psychol.* **39**, 913–924 (2011).
72. A. D. Pickering, F. Pesola, Modeling dopaminergic and other processes involved in learning from reward prediction error: Contributions from an individual differences perspective. *Front. Hum. Neurosci.* **8**, 740 (2014).
73. L. D. Smillie, A. J. Cooper, A. D. Pickering, Individual differences in reward–prediction–error: Extraversion and feedback-related negativity. *Soc. Cogn. Affect. Neurosci.* **6**, 646–652 (2011).
74. E. H. Patzelt, C. A. Hartley, S. J. Gershman, Computational phenotyping: Using models to understand individual differences in personality, development, and mental illness. *Pers. Neurosci.* **1**, E18 (2018).
75. B. F. Jeronimus, R. Kotov, H. Riese, J. Ormel, Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: A meta-analysis on 59 longitudinal/prospective studies with 443 313 participants. *Psychol. Med.* **46**, 2883–2906 (2016).
76. B. L. Goldstein, R. Kotov, G. Perlman, D. Watson, D. N. Klein, Trait and facet-level predictors of first-onset depressive and anxiety disorders in a community sample of adolescent girls. *Psychol. Med.* **48**, 1282–1290 (2018).
77. L. A. Clark, D. Watson, S. Mineka, Temperament, personality, and the mood and anxiety disorders. *J. Abnorm. Psychol.* **103**, 103–116 (1994).
78. R. Kotov, W. Gamez, F. Schmidt, D. Watson, Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychol. Bull.* **136**, 768–821 (2010).
79. D. N. Klein, R. Kotov, S. J. Bufferd, Personality and depression: Explanatory models and review of the evidence. *Annu. Rev. Clin. Psychol.* **7**, 269–295 (2011).
80. J. M. Hettema, M. C. Neale, J. M. Myers, C. A. Prescott, K. S. Kendler, A population-based twin study of the relationship between neuroticism and internalizing disorders. *Am. J. Psychiatry* **163**, 857–864 (2006).
81. A. Caspi, B. W. Roberts, R. L. Shiner, Personality development: Stability and change. *Annu. Rev. Psychol.* **56**, 453–484 (2005).
82. D. H. Barlow, K. K. Ellard, S. Sauer-Zavala, J. R. Bullis, J. R. Carl, The origins of neuroticism. *Perspect. Psychol. Sci.* **9**, 481–496 (2014).
83. J. A. Clauss, J. U. Blackford, Behavioral inhibition and risk for developing social anxiety disorder: A meta-analytic study. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 1066–1075.e1 (2012).
84. C. C. Conway, M. G. Craske, R. E. Zinbarg, S. Mineka, Pathological personality traits and the naturalistic course of internalizing disorders among high-risk young adults. *Depress. Anxiety* **33**, 84–93 (2016).
85. S. Lissek, D. E. Bradford, R. P. Alvarez, P. Burton, T. Espensen-Sturges, R. C. Reynolds, C. Grillon, Neural substrates of classically conditioned fear-generalization in humans: A parametric fMRI study. *Soc. Cogn. Affect. Neurosci.* **9**, 1134–1142 (2014).
86. O. Laufer, D. Israeli, R. Paz, Behavioral and neural mechanisms of overgeneralization in anxiety. *Curr. Biol.* **26**, 713–722 (2016).
87. B. W. Feather, Semantic generalization of classically conditioned responses: A review. *Psychol. Bull.* **63**, 425–441 (1965).
88. J. Rogers, S. Shelton, W. Shelledy, R. Garcia, N. Kalin, Genetic influences on behavioral inhibition and anxiety in juvenile rhesus macaques. *Genes, Brain Behav.* **7**, 463–469 (2008).
89. D. R. Hirshfeld, J. F. Rosenbaum, J. Biederman, E. A. Bolduc, S. V. Faraone, N. Snidman, J. S. Reznick, J. Kagan, Stable behavioral inhibition and its association with anxiety disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **31**, 103–111 (1992).
90. M. Botvinick, A. Weinstein, A. Solway, A. Barto, Reinforcement learning, efficient coding, and the statistics of natural tasks. *Curr. Opin. Behav. Sci.* **5**, 71–77 (2015).
91. S. G. Shamay-Tsoory, A. Mendelsohn, Real-life neuroscience: An ecological approach to brain and behavior research. *Perspect. Psychol. Sci.* **14**, 841–859 (2019).
92. A. R. Otto, A. Skatova, S. Madlon-Kay, N. D. Daw, Cognitive control predicts use of model-based reinforcement learning. *J. Cogn. Neurosci.* **27**, 319–333 (2014).
93. J. Alexander Jr., T. E. Audesirk, G. J. Audesirk, One-trial reward learning in the snail *Lymnea stagnalis*. *J. Neurobiol.* **15**, 67–72 (1984).
94. H. J. Eysenck, Single-trial conditioning, neurosis, and the Napalkov phenomenon. *Behav. Res. Ther.* **5**, 63–65 (1967).
95. A. R. Otto, S. M. Fleming, P. W. Glimcher, Unexpected but incidental positive outcomes predict real-world gambling. *Psychol. Sci.* **27**, 299–311 (2016).
96. A. R. Otto, J. C. Eichstaedt, Real-world unexpected outcomes predict city-level mood states and risk-taking behavior. *PLOS ONE* **13**, e0206923 (2018).
97. C. Gagne, P. Dayan, S. J. Bishop, When planning to survive goes wrong: Predicting the future and replaying the past in anxiety and PTSD. *Curr. Opin. Behav. Sci.* **24**, 89–95 (2018).
98. Q. J. Huys, R. Cools, M. Gölzer, E. Friedel, A. Heinz, R. J. Dolan, P. Dayan, Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLOS Comput. Biol.* **7**, e1002028 (2011).
99. H. J. Eysenck, Neuroticism, anxiety, and depression. *Psychol. Inq..* **2**, 75–76 (1991).
100. R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **166**, 1092–1097 (2006).
101. K. Friston, J. Daunizeau, S. Kiebel, Reinforcement learning or active inference? *PLOS ONE* **4**, e6421 (2009).
102. A. S. Heller, From conditioning to emotion: Translating animal models of learning to human psychopathology. *Neuroscientist* **26**, 43–56 (2020).
103. W. A. Hershberger, An approach through the looking-glass. *Anim. Learn. Behav.* **14**, 443–451 (1986).
104. E. Schulz, R. Bhui, B. C. Love, B. Brier, M. T. Todd, S. J. Gershman, Structured, uncertainty-driven exploration in real-world consumer choice. *Proc. Natl. Acad. Sci.* **116**, 13903–13908 (2019).
105. T. L. Griffiths, Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015).
106. R. F. Baumeister, E. Bratslavsky, C. Finkenauer, K. D. Vohs, Bad is stronger than good. *Rev. Gen. Psychol.* **5**, 323–370 (2001).
107. L. A. McCloskey, M. Walker, Posttraumatic stress in children exposed to family violence and single-event trauma. *J. Am. Acad. Child Adolesc. Psychiatry* **39**, 108–115 (2000).
108. E. Vrieze, D. A. Pizzagalli, K. Demyttenaere, T. Hompes, P. Sienaert, P. de Boer, M. Schmidt, S. Claes, Reduced reward learning predicts outcome in major depressive disorder. *Biol. Psychiatry* **73**, 639–645 (2013).
109. M. F. Scheier, C. S. Carver, M. W. Bridges, Optimism, pessimism, and psychological well-being, in *Optimism & Pessimism: Implications for Theory, Research, and Practice* (American Psychological Association, 2001), pp. 189–216.
110. C. S. Carver, Optimism and pessimism, in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (Elsevier Inc., 2015), pp. 263–267.
111. D. G. Williams, Dispositional optimism, neuroticism, and extraversion. *Personal. Individ. Differ.* **13**, 475–477 (1992).

112. S. Mineka, J. F. Kihlstrom, Unpredictable and uncontrollable events: A new perspective on experimental neurosis. *J. Abnorm. Psychol.* **87**, 256–271 (1978).
113. H. Yanagisawa, O. Kawamata, K. Ueda, Modeling emotions associated with novelty at variable uncertainty levels: A Bayesian approach. *Front. Comput. Neurosci.* **13**, 2 (2019).
114. M. Zuckerman, Diathesis-stress models, in *Vulnerability to Psychopathology: A Biosocial Model* (American Psychological Association, 1999), pp. 3–23.
115. Qualtrics, Qualtrics (2020); www.qualtrics.com.
116. C. J. Soto, O. P. John, Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *J. Res. Pers.* **68**, 69–81 (2017).
117. R Core Team, R: A language and environment for statistical computing (2017); www.R-project.org/.
118. D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, G. Grothendieck, P. Green, lme4: Linear mixed-effects models using "Eigen" and S4 (2018); https://CRAN.R-project.org/package=lme4.
119. P.-C. Bürkner, brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
120. S. J. Gershman, The generative adversarial brain. *Front. Artif. Intell.* **2**, 18 (2019).

Acknowledgments

Funding: This work was funded by the National Institute of Mental Health of the National Institutes of Health grant R21MH125311 (to A.S.H.). **Author contributions:** Conceptualization: A.S.H. and W.J.V. Methodology: A.S.H., W.J.V., N.I.K., T.R.R., B.A.J., and A.R.O. Investigation: A.S.H., W.J.V., N.I.K., T.R.R., and B.A.J. Visualization: W.J.V. Supervision: A.S.H. and A.R.O. Writing—original draft: W.J.V. and A.S.H. Writing—review and editing: A.S.H., W.J.V., N.I.K., T.R.R., B.A.J., and A.R.O. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Deidentified data and analysis code are available on the Open Science Framework (https://osf.io/fasr9/).

Submitted 2 June 2022
Accepted 30 November 2022
Published 4 January 2023
10.1126/sciadv.add2976