

**LEARNING TO FEAR SOCIAL INTERACTIONS:  
DYSREGULATED NEURAL MECHANISMS OF SOCIAL LEARNING  
IN ADOLESCENT SOCIAL ANXIETY**

---

A Dissertation  
Submitted to  
the Temple University Graduate Board

---

In Partial Fulfillment  
of the Requirements for the Degree  
DOCTOR OF PHILOSOPHY

---

by  
Tessa Clarkson  
August 2023

Examining Committee Members:

Dr. Johanna Jarcho, Advisory Chair, Department of Psychology  
Dr. Tom Olino, Committee Chair, Department of Psychology  
Dr. Phil Kendall, Department of Psychology  
Dr. Jason Chien, Department of Psychology  
Dr. Leor Hackel, External Member, University of Southern California  
Dr. Matthew Lerner, External Member, Stony Brook University

## ABSTRACT

Social anxiety (SA) disorder is prevalent, chronic, and impairs quality of life. Typical onset occurs in early adolescence, when social relationships become more salient and complex. Difficulty learning from these newly complex, important interactions may potentiate SA. SA is associated with suboptimal adaptive learning rates in non-social and uncertain contexts. However, the impact of social contexts on learning during social interactions with peers in SA remains unknown. Enhanced SA symptoms while anticipating social feedback is associated with dysregulated engagement of neural circuits implicated in salience and reward processing, which are critical hubs for learning. Despite this overlap, the neural mechanisms that support learning from social feedback remain relatively unexplored in SA. To study the influence of social context and feedback on learning in SA, we paired computational modeling with a novel social interaction fMRI task to determine the extent to which peer value as well as the valence and predictability of peer feedback modulate the neural bases of social learning about peers and their relation to adolescent SA. Fifty-nine youth (age 10-15yrs) with a range of SA engaged in real-time social interactions with purported peers while undergoing an fMRI scan. Youth with clinically relevant SA learn to fear social interactions by emphasizing unexpected negative feedback, and discounting unexpected positive feedback, in predictably nice and unpredictable social contexts to rapidly adjust learning. More severe SA was associated with decreased engagement of salience, reward and cognitive control regions in predictably nice social contexts, and only engaged cognitive control regions in

unpredictably mean social contexts. Results elucidate which aspects of the social context contribute to learning to fear social interactions on a neural and behavioral level that potentiate SA, which can inform specific intervention targets.

This dissertation is dedicated to my  
family, friends, and mentors  
who have supported me throughout.

## **ACKNOWLEDGMENTS**

I would like to thank Leor Hackel and David Barack for supporting me in learning computational modeling. I would also like to thank Megan Quarmley, Camillie Johnston, and Alisha Arora for working through the COVID pandemic to help with data collection and recruitment in order to complete this study. I would also like to thank Johanna Jarcho for her mentorship and guidance throughout this project in order to make it happen. Furthermore, this study was supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number F31MH122091. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Tessa Clarkson was also supported by the Temple University Dissertation Completion Grant, the Temple University Public Policy Lab Graduate Fellowship, the American Psychological Association (APA) Dissertation Research Award, and the Dr. Phillip J Bersh Memorial Student Award.

# TABLE OF CONTENTS

	Page
ABSTRACT.....	ii
DEDICATION.....	iv
ACKNOWLEDGMENTS.....	v
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x
CHAPTER 1	
1. INTRODUCTION.....	1
Adolescent Susceptibility for Social Anxiety.....	2
Computational Models of Learning and Their Neural Correlates.....	3
Extending Learning Models to Social Anxiety.....	4
Neural Mechanisms that Support Social Learning and Social Anxiety	
Overlap.....	5
Dysregulated Brain Function is Associated with Differences in Learning	
in Social Anxiety.....	6
Current Study.....	7
2. METHODS.....	9
Participants.....	9
Anxiety Assessment.....	10
Anxiety Diagnostic Inventory Schedule.....	11
Screen for Child Anxiety Related Emotional Disorders.....	11

LEARN Task.....	12
Task Composition .....	13
Social Context Factors.....	14
Debriefing.....	15
fMRI Data Acquisition .....	16
Data Analyses.....	17
Computational Models .....	17
Model Fitting & Comparison.....	24
Model Parsimony Analyses .....	25
Simulation: Model-Based Regressor Generation.....	26
Model-Based fMRI Analyses .....	28
3. RESULTS .....	33
Model Comparison .....	33
Model Validation of M10 $\lambda_{pos/neg}$ .....	34
Learning Analyses.....	34
Summary of Computational Modeling and Learning Analyses. ....	35
Model-Based fMRI Analyses.....	37
Reputation-Based Prediction Error: Valence by Predictability by	
SA.....	37
Accuracy-Based Prediction Error: Valence by Predictability by	
SA.....	38
4. DISCUSSION .....	39
Mechanisms of Social Learning in Youth.....	39

Distinct Mechanisms of Social Learning Social Anxiety .....	40
Model-Based Prediction Error Modulated Neural Activation .....	43
Neural Mechanisms of Social Learning in Social Anxiety .....	44
Conclusions .....	45
Limitations & Future Directions .....	46
BIBLIOGRAPHY .....	47

## LIST OF TABLES

Table	Page
1. Demographics.....	10
2. Bayesian Model Comparison.....	33
3. Model Validation: Average Prediction Proportion by Social Context and SA...	34
4. Group-level Activation Clusters for Reputation-Based Prediction Error: Valence by Predictability by SA. ....	38

## LIST OF FIGURES

Figure	Page
1. <u>LEARN Task</u> . Each social interaction trial includes portions A-C followed by an inter-trial interval. ....	13
2. <u>Apriori Network Mask Sub-Regions</u> . Bilateral- A) vmPFC, B) anterior dACC, C) posterior dACC, D) anterior insula, E) ventral striatum .....	31
3. <u>Model Validation of M10 <math>\lambda</math> pos/neg</u> . Compares youth with clinically significant SA ( $\geq 7$ ) vs subclinical or no SA ( $< 7$ ) on actual and simulated learning in the last 8 trials across social contexts. The x-axis show the predicted valence, y-axis is the average proportion of the predicted choice. ....	34
4. <u>Parameters of Social Learning Across Contexts and SA in Youth</u> . Compares youth with clinically significant SA ( $\geq 7$ ) vs subclinical or no SA ( $< 7$ ) on parameter estimates of A) learning rates and B) associative values across social contexts. ....	35
5. <u>Significant Clusters Reputation-Based Prediction Error: Valence by Predictability by SA</u> . A) bilateral vmPFC, B) right vmPFC, C) Left Anterior dACC, D) Right Posterior dACC, E) Left Anterior Insula F) Left Ventral Striatum, G) Right Ventral Striatum.....	37
6. <u>Decomposition of Reputation-Based Prediction Error: Valence by Predictability by SA within Activation Clusters</u> . The x-axis is social anxiety centered at 7, thus values on the right side of the dotted line represent youth with clinically significant SA and values on the left side of the dotted line represent youth without SA. The y-axis in the effect size of the reputation-based prediction error on the neural signal in each region, with larger numbers indicating a greater increases or decreases in brain activation. The graph on the left-side show the relations between valence and SA while receiving feedback from unpredictable peers (Mean60-blue or Nice60-pink), and graphs on the right-side show relation these relations while receiving feedback from predictable peers (Mean80-blue or Nice80-pink). Black marks on the x-axis represent participant data range. * indicate a significant valence by SA interaction, bars indicate a significant difference in activation by SA within valence. ....	37

## CHAPTER 1

### INTRODUCTION

Social anxiety (SA) disorder is a chronic condition (Beesdo-Baum et al., 2012) that affects over 13% of the population (Kessler, 2003), and up to 36% of youth (Jefferies & Ungar, 2020). Typical onset occurs in early adolescence (Beesdo-Baum et al., 2012) and peaks at 10-15 years of age (DeWit et al., 2005), when complexity of the social milieu and desire for social acceptance increase (Bradford Brown, Eichert, & Petriet, 1986). In the context of this increasing complexity, deficits in the capacity to learn about others may contribute to a core symptom of SA: fear of negative social feedback (Boelen & Reijntjes, 2009; Caouette et al., 2015). Adaptive learning in different social contexts, that vary in valence and predictability, may be crucial for youth to adjust to rapidly changing adolescent social interactions (Crone & Dahl, 2012). Because early onset SA is associated with more severe (Beesdo-Baum et al., 2012; Bruce, Yonkers, Otto, & Eisen, 2005; Reilly-Harrington & Sachs, 2006), costly (Hidalgo, Barnett, & Davidson, 2001), comorbid (Beesdo et al., 2007) and long-lasting symptoms (Boer, 1997), isolating etopathogenic mechanisms of SA is imperative (Heiser, Turner, & Beidel, 2003; Kessler, Chiu, Demler, Merikangas, & Walters, 2005; Reilly-Harrington & Sachs, 2006). While available SA treatment can reduce symptoms, they rarely result in full remission (Ipser, Stein, Hawkrige, & Hoppe, 2009; James, Reardon, Soler, James, & Creswell, 2020). Targeting dysregulated neural activation in regions associated with social learning may enhance treatment efficacy in SA (Orr & Moscovitch, 2010; Piray, Ly, Roelofs, Cools, & Toni, 2019; Ressler, 2007). Therefore, it is critical that we investigate how learning in different social contexts modulates neural activation in SA.

## **Adolescent Susceptibility for Social Anxiety**

SA disorder is characterized by an intense fear of negative social feedback (American Psychiatric Association. & American Psychiatric Association. DSM-5 Task Force., 2013). Symptoms are potentiated in uncertain or novel contexts (Boelen & Reijntjes, 2009; Boelen, Vrinssen, & van Tulder, 2010; Carleton, Collimore, & Asmundson, 2010; Jarcho et al., 2016, 2015) and while receiving positive or negative feedback from peers (Guyer et al., 2014; Guyer, McClure-Tone, Shiffrin, Pine, & Nelson, 2009). Adolescence is a sensitive period for developing SA, likely due to an increased desire for peer acceptance (Bradford Brown et al., 1986) and the emergence of complex social relationships (Nelson & Guyer, 2011) that necessitate learning from more nuanced interactions. These changes coincide with fine-tuning of neural circuits implicated in reward and salience processing, which support social learning, decision making, and mentalizing—a component of social learning. Development of brain regions associated with social learning (ventral medial prefrontal cortex (vmPFC)), insula, and dorsal anterior cingulate cortex (dACC) during adolescence are related to changes in neural responses to social experiences (Anderson, Bechara, Damasio, Tranel, & Damasio, 1999; Boehme et al., 2017; Monahan, Guyer, Silk, Fitzwater, & Steinberg, 2016; Raznahan et al., 2010) and peer feedback (Guyer, Choate, Pine, & Nelson, 2012). Difficulty learning about others may promote fear of negative evaluation, a hallmark symptom of SA disorder (Morrison & Heimberg, 2013). Thus, early adolescence is a critical age for isolating maladaptive mechanisms of social learning, which may lay the foundation for severe and chronic SA.

## **Computational Models of Learning and Their Neural Correlates**

Well-established computational models use different parameters to estimate learning in the non-social domain. Reinforcement learning models estimate learning by computing prediction errors, such that positive errors reflect outcomes that were better than expected and negative errors reflect outcomes that were worse than expected. Learning is achieved by updating expectations in proportion to the prediction error and a constant learning rate (Rescorla & Wagner, 1972.). Alternatively, associative learning models estimate learning via prediction errors, but adjust learning rates over time based on the degree of surprise about an outcome, which allows learning to speed up after more surprising outcomes and slow down with less surprising outcomes, thereby fine tuning learning over time (Pearce, Kaye, & Hall, 1982). Recent studies have combined these approaches to allow learning rates to be dynamically adjusted based on the valence of the prediction error, surprise of the outcome, and level of predictability of the context (Behrens, Hunt, Woolrich, & Rushworth, 2008; Behrens, Woolrich, Walton, & Rushworth, 2007; Li et al., 2011; Piray et al., 2019).

Reinforcement and associative learning are encoded in overlapping brain regions shared by (vmPFC) and unique to the reward (ventral striatum) and salience (anterior insula and dACC) networks. In the reward network, the ventral striatum encodes the perceived positive or negative value of a stimulus (Murray & Wise, 2010; Padoa-Schioppa & Assad, 2006; Schoenbaum, Roesch, Stalnaker, & Takahashi, 2009) and varying degrees of perceived predictability of the reward (R. M. Jones et al., 2011; Poore et al., 2012). Specifically, larger positive prediction-errors while receiving feedback, evince greater ventral striatal activity, whereas larger negative prediction-errors evince

decreased striatal activation (Abler, Walter, Erk, Kammerer, & Spitzer, 2006; Bhanji & Delgado, 2014; Harris & Fiske, 2010; Yin et al., 2009). Similarly, in the salience network, the dACC and anterior insula encode the perceived positive or negative value of a stimulus (Behrens et al., 2007; Kuhnen & Knutson, 2005; Padilla-Coreano, Tye, & Zelikowsky, 2022), as well as varying degrees of perceived *predictability* in the context (Behrens et al., 2007; Piray et al., 2019). Specifically, greater environmental unpredictability is associated with greater insula and dACC activity. Further evidence linking the dACC and learning shows that manipulations of dACC activity during positive and negative feedback scenarios alters learning processes (Klöbl et al., 2020). Moreover, an overlapping region of the reward and salience network, the vmPFC (Bolstad et al., 2013; Ernst & Fudge, 2009; Groenwegen, Vermeulen-Van Der Zee, Kortschot, & Wittex, 1987), is critical for learning based on past experiences (Behrens et al., 2009, 2008; Bhanji & Delgado, 2014; Padilla-Coreano et al., 2022). Thus, learning may be best understood by probing regions associated in both the reward and salience networks.

### **Extending Learning Models to Social Anxiety**

Although primarily implemented in non-social domains, the parameters of computational learning models can be leveraged to isolate the etopathogenic mechanisms of SA. Learning to adapt to nuanced social situations is critical to social competence. Indeed, SA is associated with deficits in learning and recalling social rules (Lewis et al., 2017), updating predictions about social feedback (Jarcho et al., 2015), and maladaptive behavioral responses to that feedback. Although limited research has probed the neural mechanisms of social learning in SA, emerging evidence suggests parameters

critical to learning elicit altered neural response in the reward and salience networks during social feedback with unpredictable peers (Bradford Brown et al., 1986; Jarcho et al., 2015), particularly when they provide positive feedback (Boelen & Reijntjes, 2009; Carleton et al., 2010; Jarcho et al., 2015). Yet the extent to which seemingly critical parameters of the social context influence peer-based learning are largely unknown. Therefore, examining the neural and behavioral mechanisms of that may influence symptoms trajectories SA is critical.

### **Neural Mechanisms that Support Social Learning and Social Anxiety**

#### **Overlap**

As in non-social learning, reward and salience networks are also implicated in social learning. In adolescents, faster social learning evinces stronger within reward-network coupling during negative-vs-positive feedback (Van Den Bos, Cohen, Kahnt, & Crone, 2012). Activation within the salience network also varies as a function of the valence of peer feedback (Jarcho et al., 2016), which may be a critical parameter of social learning. In social contexts, aspects of the reward network (ventral striatum) may be involved in social reward processing and reinforcement learning while aspects of the salience network (dACC, anterior insula) may enhance reward learning by responding to the degree of surprise of the social feedback in associative learning during social interactions (Feng et al., 2021; Fouragnan, Retzler, & Philiastides, 2018; Garrison, Erdeniz, & Done, 2013a; Lindström, Haaker, & Olsson, 2018a; Veit et al., 2012). These networks may function together to facilitate social learning (Franco Cauda, Andrea E. Cavanna, Federico D'agata, Katuscia Sacco, Sergio Duca, 2011; Hollmann et al., 2011). Thus, maladaptive responses in either network and/or shared cognitive control hubs, such

as the vmPFC (Behrens et al., 2008; Feng et al., 2021; Padilla-Coreano et al., 2022), may underlie neurobiological susceptibility to SA in adolescence. Indeed, social interactions elicit dysregulated engagement in overlapping network hubs in SA (Caouette & Guyer, 2014; Freitas-Ferrari et al., 2010; Jarcho et al., 2016; Yang et al., 2017). We demonstrated that dysregulated neural responses in salience and reward networks during unpredictably positive social feedback in preadolescents at risk for SA disorder predicted more severe SA symptoms in mid-adolescence (Clarkson et al., 2019). Thus, both social learning and SA are associated with highly convergent neural mechanisms.

### **Dysregulated Brain Function is Associated with Differences in Learning in Social Anxiety**

Few studies have investigated neural correlates of learning in adolescents with SA. Our previous work examined neural correlates of prediction-errors to social feedback in adolescents with SA. We found prediction errors to unexpected positive feedback from peers elicited heightened reward network activity and negative functional connectivity within the reward, salience, and overlapping network regions in anxious-vs-healthy adolescents (Jarcho et al., 2015). The degree to which these regions exhibited negative functional connectivity predicted impaired recall for unexpected positive social feedback (Jarcho et al., 2015). These findings suggest that differences in neural response to positive prediction-errors may negatively bias future expectations about social feedback in SA (Morrison & Heimberg, 2013). However, this study did not use model-derived prediction errors, and therefore it is difficult to make generalizations regarding the mechanisms of social learning in SA. Another study examined relations of reinforcement, associative, and combined learning models during non-social learning and brain function

and found that a combined model best captured learning rates, such that impaired dynamic learning of emotionally salient stimuli in SA was associated with dysregulated activation in the salience network (Piray et al., 2019). These findings suggest that individuals with SA adjust their learning over time, and the degree of associability (i.e., attention allocated to the prediction error vs their already established expectations) was encoded in the salience network activation in SA. Yet, this study did not examine learning in a social context, which typically engages reward network regions given the rewarding nature of social feedback.

### **Current Study**

The present study addresses prior limitations by using computational modeling and functional neuroimaging to investigate mechanisms of social learning that relate to SA in adolescents. We compared reinforcement learning models, associative value models, and combined models in social contexts that vary in their predictability and valence to understand differences in social learning among adolescents with a full range of SA symptoms. Using the fMRI-based Learning from Evaluation And Recall of interactions (LEARN) task, a novel variant of the Virtual School Task (Clarkson et al., 2019; Clarkson, Karvay, Quarmley, & Jarcho, 2021; Jarcho et al., 2016; Smith et al., 2020), we manipulated the social context by adjusting the valence and predictability of purported peer feedback and allowing youth to learn about their peers over iterative social interactions. We then estimated learning by comparing computational models that prioritize different aspects of the social context (i.e. valence and predictability) as well as how learning may change over time based on the associability placed on one's existing knowledge of a peer's reputation vs. the surprise of the feedback from the last trial. We

used a full range of SA symptoms as youth with subclinical levels of SA may later develop clinically relevant symptoms. Winning model parameters were compared across adolescents with a full range of SA symptoms to determine differences in the mechanisms of social learning in SA. Model parameters were then used in a fMRI analyses to elucidate how differences in social learning modulate neural activation across levels of SA. Behaviorally, we predicted that individuals with more severe SA would differ in learning across social contexts and that this would change rapidly over time based on the associability placed on one's existing knowledge of a peer's reputation. Specifically, we predicted that youth with SA would have difficulty learning from positive social feedback (Jarcho et al., 2015; Quarmley, Nelson, Clarkson, White, & Jarcho, 2019), and unpredictable social contexts (Clarkson et al., 2019; Jarcho et al., 2016), and that their learning would adjust rapidly based on the degree of the surprise of the feedback from the last trial (Piray et al., 2019). In terms of brain function, we predicted that model-based prediction-errors would be associated with increases in brain activation across the salience and reward networks, specifically the vmPFC, dACC, anterior insula, and ventral striatum (Feng et al., 2021; Padilla-Coreano et al., 2022), and that relations would vary based on the social contexts (i.e. valence and predictability).

## METHODS

### Participants

Participants (10-15 years of age; Table 1) were recruited from the greater Philadelphia metro area and from the Child and Adolescent Anxiety Disorders Clinic (CAADC) at Temple University. During a phone interview, potential participants were screened for exclusionary criteria (e.g. contraindications for fMRI, neurological, or physical impairment, primary mental health diagnoses other than anxiety, head injuries with loss of consciousness or unstable medication < 2months). Participants completed 2-4 study visits, 1-3 of which were completed virtually due to the COVID-19 pandemic, and the final fMRI-visit occurred in-person. The first two virtual visits included parent and child structured diagnostic interviews, which were either done by the reliable clinical psychology graduate students, or skipped if participants previously completed these interviews as part of the intake process at the CAADC clinic before enrolling in the study. The third virtual and fourth in-person visits were completed by all participants, which included questionnaires and a behavioral task in visit 3 to setup the visit 4 fMRI task. All visits were completed within 4 weeks from the initial visit. All participants and their parents consented/assented to study procedures, which were approved by the Temple University Internal Review Board.

Fifty-nine youth enrolled in the study. Four participants withdrew (N=4). The remaining, 47 were included in computational modeling analyses. Participants were excluded for technical issues that resulted in behavioral data loss (N=4), >40% missed responses (N=2), or response selections of only one option (N=2). Of those 47 participants, 33 were also included in the model-based neuroimaging analyses. These participants were excluded for technical issues with the scan visit (i.e. pulse trigger issue,

incomplete scan, internal communication errors on scanner, stimulus computer crashing; N=9) and for motion or artifacts (N=5). Participants included in the computational modeling subsample and the model-based neuroimaging subsample did not differ by age, gender, family income, or race ( $p$ 's>0.319; Table 1).

**Table 1. Demographics**

	<b>Computational Modeling Sample (N=47)</b>	<b>Neuroimaging Sample (N=33)</b>	<b>p-value</b>
<b>Age Mean (SD)</b>	12.53 (1.42)	12.73 (1.38)	0.319
<b>Gender Ratio (M:F)</b>	26:21	18:15	1.000
<b>Mean Family Income (range)</b>	\$71,867 (\$0-\$285,000)	\$76,710 (\$0-\$285,000)	0.747
<b>Race</b>			0.791
<b>White</b>	33	22	
<b>Multi-Racial</b>	8	6	
<b>Black/African American</b>	4	3	
<b>Asian/Pacific Islander</b>	2	2	
<b>Anxiety Diagnosis: ADIS</b>	17	14	0.469
<b>Social Anxiety: SCARED</b>	4.89 (3.97)	5.00 (4.24)	0.906

### **Anxiety Assessment**

In the present study we used the dimensional value of SA in all analyses from the the Screen for Child Anxiety Related Emotional Disorders (SCARED; Muris, Merckelbach, Schmidt, Mayer, & Birgit, 1999) to better understand behavioral and neural mechanisms of social learning in adolescent SA. However, we also checked for correspondence between child-reported dimensional symptoms on the SCARED and ratings for clinical psychology graduate students on the parent and child Anxiety Diagnostic Inventory Schedule (ADIS-5; Brown & Barlow, 2014) interviews to ensure accurate representation of SA. Decomposition analyses split the sample based on a clinical-cut of score that was consistent with ADIS diagnoses for visualization.

### ***Anxiety Diagnostic Inventory Schedule.***

On the first visit, parents completed the parent consent form and the parent Anxiety Diagnostic Inventory Schedule (ADIS-5; Brown & Barlow, 2014) interview virtually. On the second visit, youth completed the child assent form and child ADIS interview virtually. Participants recruited from the CAADC complete the ADIS interviews during the CAADC intake process. Therefore, their CAADC intake interview results were used in lieu of completing visits 1 & 2, and consent/assent was completed at “visit 3”. Using combined information from the parent and child ADIS interviews, N=17 youth met criteria for a clinical diagnosis (Clinical Severity Rating  $\geq 4$  on either/both parent or child ratings) of an anxiety disorder in the computational modeling sample (36%) and 14 met criteria in the neuroimaging subsample (42%). There were no significant differences between proportion of clinically anxious youth between the samples ( $p=0.469$ ; Table 1).

### ***Screen for Child Anxiety Related Emotional Disorders.***

On the third visit, participants completed questionnaires; research staff regularly checked in to ensure comprehension and proper follow-up to any endorsed questions regarding child abuse or suicidality, in line with mandated reporting guidelines. One of these questionnaires was the Screen for Child Anxiety Related Emotional Disorders (SCARED; Muris, Merckelbach, Schmidt, Mayer, & Birgit, 1999), which is a child-self report measures that contains five reliable ( $\alpha = 0.90$ ) and valid (Birmaher et al., 1999) subscales including social, school phobia, generalized anxiety, separation anxiety, panic, and total anxiety symptoms. Higher scores indicate more severe symptoms. Twelve

youth (26%) endorsed clinically relevant social anxiety symptoms ( $\geq 8$ ) and seven (15%) endorsed sub-clinical social anxiety symptoms (5-7) in the computational modeling sample (N=47). Nine youth (27%) endorsed clinically relevant social anxiety symptoms ( $\geq 8$ ) and five (15%) endorsed sub-clinical social anxiety symptoms (5-7) in the neuroimaging sample. There were no differences average social anxiety scores between the samples ( $p=0.906$ ; Table 1).

Participants endorsing clinically relevant social anxiety symptoms ( $\geq 8$ ) or elevated sub-clinical symptoms (7) on the SCARED, also met diagnostic criteria on the ADIS in both subsamples. Because continuous measures of anxiety allow for better understanding individual differences and how social anxiety may act dimensionally, and continuous thresholds corresponded with diagnostic cut-offs, both analyses used the continuous measure of the SCARED self-report measure to examine association of the neural basis of social learning and SA. For decomposition analyses and visualizations, the SCARED was centered at 7 to clarify interpretations between clinically significant social anxiety symptoms ( $\geq 7$ ) and subclinical or no anxiety symptoms ( $< 7$ ) and facilitate decompositions.

### **LEARN Task**

The LEARN Task (programmed in Presentation Version 23.0; Neurobehavioral Systems, Inc., Berkeley, CA, [www.neurobs.com](http://www.neurobs.com)) is a modified version of the Virtual School Task (Clarkson et al., 2021; Jarcho et al., 2016, 2019; Quarmley et al., 2019) in which participants interact with purported peers with distinct personalities. Unlike the standard Virtual School Task where participants learn the reputations of purported peers

prior to their interactions, the LEARN variant was designed to quantify experiential learning about purported peers over the course of numerous interactions.

### ***Task Composition***

The LEARN Task measured how individuals learn the reputations of different peers over the course of iterative social interactions by allowing participants to make predictions about the valence of social feedback they receive, prior to the social interaction. A complete social interaction included a prediction about the social feedback, presentation of social feedback, and the opportunity to respond to that feedback.

In preparation for the LEARN Task, on visit 3 participants created an avatar of themselves using Avatoon and made a personal profile in which they answered multiple choice questions about things they enjoy. They were told their responses would be sent to the other peers they were going to be interacting with during the LEARN Task in the scanner. At visit 4, participants completed the 26 minute-long LEARN task while undergoing an fMRI scan. During the LEARN task (Figure 1), participants entered 4 classrooms populated by 4 purported peers. Each classroom constituted 1 run, which lasted 6 minutes and 30 seconds. Each peer interacted with the participant 8 times in each of the 4 runs. This amounted to 32 interactions per peer and 128 interactions in total. The order of interactions was held-constant so that each participant had the same social experience and opportunity to learn about their peers in the same order and predict mean or nice peer feedback. This helped reduce variability in learning rates and enhanced model fitting.

Participants were told they would be interacting with peers in a virtual school and that they would have a chance to guess what they thought each peer might say before the

interaction and that they would have a chance to respond to what the peer said. For each trial, the participant predicted the valence of the social feedback they were about to receive (i.e. mean/nice) from a specific peer who had a typing bubble displayed over their avatar (Figure 1A; 4 sec). Next, the participant received social feedback from that peer (Figure 1B; 3 sec), which confirmed or disconfirmed their prediction, and resulted in a positive or negative prediction error. Then, participants made a response to complete the social interaction (Figure 1C; 4 sec). Participants made responses to feedback by selecting either, “You’re Right”, “You’re Wrong”, “You’re Nice”, or “You’re Mean” to reply to the peer. The feedback provided by each peer interaction allowed for participants to learn over time how to better predict the type of feedback they will receive from each peer during subsequent interactions. Each trial was followed by an inter-trial interval (.5 sec). Reaction times for prediction (M = 1.570, SD = 0.805, Range = 1.08 to 2.147 seconds) and response (M= 1.21, SD=0.84, Range= 0.53-4.00 seconds) were used in fMRI analyses to model the duration of each epoch and introduce jitter.

### ***Social Context Factors***

The primary goal of the study was to assess how the social context and expectations about peers differentially impact learning across SA symptoms in youth. Specifically, we examined how the social context factors of *valence* and *predictability* of peer feedback and the *associability* (i.e. attention allocated to the prediction error-vs-existing expectations (Pearce et al., 1982)) one places on a peer’s learned reputation impact social learning, and if this differs with more severe SA symptoms. Thus, each purported peer was assigned a different “reputation.” Reputation was operationalized by different proportions of nice or mean feedback over the course of the task. In this way,

both contextual factors of valence and predictability of peer feedback were manipulated, resulting in four reputation types denoted as Nice80, Nice60, Mean60, and Mean80. The predictably nice peer (Nice80) provided 80% nice and 20% mean feedback. The unpredictable nice peer (Nice60) provided 60% nice and 40% mean feedback. Likewise, the unpredictable mean peer (Mean60) provided 40% nice and 60% mean feedback while the predictable mean peer (Mean80) provided 20% nice and 80% mean feedback. Note that the context of peer reputation-based valence, which dictates the overarching or majority of the feedback a participant receives, is considered separately from individual instances of feedback, which can be positive or negative, and we refer to as feedback-based valence

The contextual factor of associability one places on a peer reputation was not explicitly manipulated. Instead, associability was modeled based on the degree to which inaccurate estimations of receiving nice feedback from a peer with a set reputation (i.e. prediction error), influenced learning rates for future estimations of receiving nice feedback from a peer with a set reputation. Greater associability of prediction errors would result in greater adjustments of future estimations of receiving nice feedback from a peer with a set reputation, and faster learning rates.

### ***Debriefing***

Following the scan, participants were debriefed and paid for their time. Deception was assessed during an interview in which participants were asked a series of increasingly specific questions about their experiences during the task. This interview culminated in the examiner explicitly asking participants if they “interacted with other peers” in the virtual school, of which 96% responded “yes” (N=45) and two “no”. These

two participants were reviewed by the research team to assess deception based on reviewing remarks these participants made during the visit about the peers they interacted with and responses about how bullied they felt during the task and were determined to have been deceived and therefore were included in the analyses.

### **fMRI Data Acquisition**

Participants first completed a 15-min mock scanning session in an MRI Stimulator<sup>TM</sup> (Psychology Software Tools, Inc), to decrease scanner-anxiety and habituate to the auditory and visual environment of the scanner, while MoTrak<sup>®</sup> software trained participants to limit their head motion. During the mock scan, participants practiced using a button box to make behavioral responses during several trials of a purportedly ‘pre-recorded’ session of the LEARN task. Participants were explicitly told that this session was recorded from another participant; therefore, they would not be interacting with real peers during these trials. To underscore the fact that practice trails were not on-line, and were not the same set of youths they would eventually interact with, depicted peers were always the opposite sex of the participant.

After mock scanning, neuroimaging data was acquired with a 3-Tesla Siemens MAGNETOM Prisma MRI (20-channel head coil). For each participant, 217 functional image volumes with 52 multiband (accel. factor 2) interleaved axial 3mm slices (in-plane resolution = 3x3mm) were acquired using a T2\*-weighted echo-planar sequence (TR/TE= 1750/29ms, flip=74°; FOV=240 mm, matrix=80x80). To facilitate anatomical localization and coregistration of functional data, a high-resolution interleaved structural scan was acquired (sagittal plane) with a T1-weighted magnetization-prepared spoiled gradient-recalled echo sequence (TE/TI= 2.17min/1150ms, flip=8°; FOV=224mm,

matrix=224x224, in-plane resolution, 1×1mm). The LEARN task was projected onto a screen and viewed via a head coil-mounted mirror. Responses were made using a fMRI-compatible button box (Current Designs, 4 button box).

## **Data Analyses**

### ***Computational Models***

**Model Comparison Summary.** Ten different computational models were compared that examined if there were different learning mechanisms based on the social context factors of interests: predictability, valence, and associability placed on one's existing knowledge of a peer's reputation. Thus, included models considered: (M1) if social contexts of reputation-based valence and predictability are irrelevant for social learning, (M2) if predictability exclusively differentiates social learning, (M3) if reputation-based valence exclusively differentiates social learning, (M4) if feedback-based valence (regardless of reputation) exclusively differentiates social learning, (M5) if predictability and reputation-based valence exclusively differentiates social learning, (M6) if associability placed on one's existing knowledge of a peer's reputation exclusively differentiates social learning, (M7) if predictability and associability placed on one's existing knowledge of a peer's reputation exclusively differentiates social learning, (M8) if reputation-based valence and associability placed on one's existing knowledge of a peer's reputation exclusively differentiates social learning, (M9) if feedback-based valence and associability placed on one's existing knowledge of a peer's reputation exclusively differentiates social learning, (M10) if predictability, reputation-based valence and associability placed on one's existing knowledge of a peer's reputation

all differentiates social learning. The winning model (see section Model Fitting & Comparison) altered for parsimony analyses (see section Model Parsimony Analyses) to determine if each parameter also was necessary and distinct for each social context.

**Shared Model Components.** All modeling was performed in R-studio version 1.4.1106 using maximum likelihood estimations from the “optim” function and finite difference approximation gradient functions. All models used the same reinforcement-learning (RL) model framework. Each model tracks the expected value of the valence of the feedback (Figure 1A)  $X_t$  (initially set to 0.5, relative to nice feedback as 1 and mean as 0) on trial  $t$  for the peer about to provide feedback and the actual prediction choice (coded: 1=nice, 0=mean) the participant made about the valence of the feedback on trial  $t$ . Each peer’s expected values were updated independently. Therefore, expected values reflected the estimated probability of receiving nice feedback from each peer. Thus, if  $S_t$ , is the peer stimulus (i.e. the peer: Mean80, Mean60, Nice60, or Nice80) about to provide feedback, and  $C_t$  is the prediction choice the participant made (i.e., 1=nice, 0=mean), and  $O_t$  is the outcome of the feedback on trial  $t$ , all models compute a prediction error signal and update the corresponding expected value:

$$\delta_t = O_t - X_t(S_t, C_t)$$

$$X_{t+1}(S_t, C_t) = X_t(S_t, C_t) + \alpha_t \delta_t$$

where  $\delta_t$  is the prediction error on trial  $t$  and  $\alpha_t$  is the learning rate that scales the influence of the prediction error on the current expected value.

Each of the models defines the learning rate differently. Models 1-5 use static learning rates, which assumes that learning happens at the same rate over the entire course of the task. In contrast, models 6-10 use dynamic learning rates, which assume

that learning rates are adjusted throughout the task and may speed up or slow down depending on learning environment. The dynamic learning rates in models 6-10 are updated based on the associative value of a particular peer to capture how much associability or attention is allocated to the prediction error vs their already established expectations of the likelihood of receiving nice feedback from a particular peer. Thus, the learning rate is adjusted over time based on how much surprise a participant is experiencing compared to what they expected. Learning rate parameters are varied for each model in order to examine the effect of reputation-based valence, predictability, and reputation associability on learning.

Additionally, each model used a choice equation to generate predictive probabilities of participants' choice data using a softmax function with a temperature ( $\tau$ ) parameter:

$$P(S_t, C_t) = \frac{1}{1 + e^{-(X_t(S_t, C_t) - (1 - X_t(S_t, C_t))) * \tau}}$$

where  $P(S_t, C_t)$  is the probability of the participant's choice for a particular stimulus (peer), and  $(X_t(S_t, C_t))$  is the expected value of the choice for a particular stimulus (peer) and  $\tau$  is the temperature, which controls the probability distribution between the expected values of each choice.

**M1. RL Non-individuation Model.** This model assumes learning happens in the same way regardless of manipulated social context factors (i.e. predictability and reputation-based valence of each peer). Thus, this model includes a single constant learning rate bounded between [0 and 1] for all peer stimulus trials.

**M2. Predictability RL Model.** This model assumes learning happens differently in predictable-vs-unpredictable social contexts, regardless of the reputation-based valence. This model ignores the valence of a peer's reputation and assigns separate

learning rates from the predictability of their reputation. Specifically,  $\alpha_{t \text{predictable}}$  for predictable peers (Nice80 and Mean80), and  $\alpha_{t \text{unpredictable}}$  for unpredictable peers (Nice60 and Mean60).

**M3. Valenced RL Model.** This model assumes learning happens differently in mean-vs-nice social contexts, regardless of the predictability of the social interaction. This model ignores the predictability of a peer's reputation and assigns separate learning rates from the reputation-based valence. Specifically,  $\alpha_{t \text{mean}}$  for mean peers (Mean60 and Mean80), and  $\alpha_{t \text{nice}}$  for nice peers (Nice60 and Nice80).

**M4. Feedback RL Model.** This model assumes learning happens in the same way regardless of the social context (i.e. predictability and reputation-based valence of each peer), however the rate of learning differs based on the valence of the feedback they receive through social interactions (feedback-based valence). This model ignores the predictability and valence of a peer's reputation and assigns separate learning rates for positive and negative feedback trials regardless of peer reputation. Specifically,  $\alpha_{t \text{neg}}$  for negative and  $\alpha_{t \text{positive}}$  for positive feedback trials (Maia & Frank, 2011).

**M5. Predictability-Valence RL Model.** This model assumes learning happens differently in mean-vs-nice and predictable-vs-unpredictable social contexts. This model assigns separate learning rates for each peer reputation. Specifically,  $\alpha_{t \text{mean80}}$  for the predictable mean peer (Mean80),  $\alpha_{t \text{mean60}}$  for the unpredictable mean peer (Mean60),  $\alpha_{t \text{nice80}}$  for predictable nice peer (Nice80), and  $\alpha_{t \text{nice60}}$  for the unpredictable nice peer (Nice60).

**M6. Associative Value-RL Model.** This model assumes learning happens in the same way regardless of the social context (i.e. predictability and valence of each peer), but instead that learning is driven by the associability or attention allocated to the prediction error vs. their already established expectations of the likelihood of receiving

nice feedback from a particular peer. Meaning, that learning rates changes dynamically over time based on how much surprise a participant is experiencing. If a participant is experiencing a lot of surprise, their current estimate may not be well calibrated and they should increase their learning rate to adapt to new feedback. If they are experiencing minimal surprise, their current estimate may be well calibrated and they should have a low learning rate that down weights new surprise feedback. To do this, the model constructs and scales the learning rate for the current trial using the associative value  $A_t$  from the previous trial (initially set to 0.5). Then the dynamic learning rate is input into the same RL model as M1-5.

$$K_t = wA_{t-1} + (1 - w)$$

$$\alpha_t = k K_t$$

First,  $K_t$  is defined above, where  $w$  is a weight parameter bounded from [0 to 1], to allow for the learning rate to be a constant, unaffected by the associative value  $A_t$  (when  $w=0$ ) or dynamically adjusted by the associative value  $A_t$ . Then  $K_t$  is scaled by the parameter  $k$  bounded from [0 to 1], which scales the learning rate  $\alpha_t$  to values between 0 and  $k$ , which allows for a maximum upper bound to the learning rate.

Next, the associative value  $A_t$  gets updated on each trial. First, the previous trial associative value  $A_{t-1}$  gets reduced gradually due to random diffusion  $\lambda$ , which is a parameter bounded from [0 to 1]. Then, after making a prediction and receiving the social feedback from a peer, the associability gets updated based on the degree of the ‘surprise’ of the social feedback using the squared prediction error multiplied by the proportion of the associability not reduced through diffusion. A larger  $\lambda$  indicates a

slower decay of the associative value and less weight of the prediction error, meaning learning is biased towards a peer's existing reputation compared to the surprise feedback.

$$A_t = \lambda A_{t-1}$$

$$A_{t+1} = A_t + (1 - \lambda)\delta_t^2$$

Thus, this model's three free parameters ( $w$ ,  $k$ ,  $\lambda$ ) are all constrained to lie in the unit range (i.e. between 0-1). Moreover, since squared prediction errors in this task are between 0 and 1, associability will also always lie in the unit range. Consequently, learning rates will always be between 0 and 1 ensuring that expected values are well-defined for any set of parameters.

**M7. Associative Value- Predictability RL Model.** This model assumes that associability or attention one places on the prediction error vs their already established expectations of the likelihood of receiving nice feedback impacts learning differently in predictable-vs-unpredictable social contexts, regardless of the valence of the reputation of the peer or of the social feedback. Meaning, how much an individual weigh's their existing conceptualization of a peer's reputation compared to the surprise feedback they just received for more predictable-vs-unpredictable peers. To do this, the model estimates two updated associative values (as in M6), one for predictable peers (Mean80 & Nice80) and one for unpredictable peers (Mean60 & Nice60) and then updates the respective learning rates in the RL, of M2. Therefore, the model estimates free parameters separately for predictable ( $w_{\text{pred}}$ ,  $k_{\text{pred}}$ ,  $\lambda_{\text{pred}}$ ) and unpredictable peers ( $w_{\text{unpred}}$ ,  $k_{\text{unpred}}$ ,  $\lambda_{\text{unpred}}$ ).

**M8. Associative Value- Valence RL Model.** This model assumes that the associability or attention one places on the prediction error vs their already established expectations of the likelihood of receiving nice feedback impacts learning differently in

mean-vs-nice social contexts, regardless of the predictability of the social interaction. Meaning, how much an individual weighs their existing conceptualization of a peer's reputation compared to the surprise feedback they just received for mean-vs-nice peers. To do this, the model estimates two updated associative values (as in M6), one for mean peers (Mean80 & Mean60) and one for nice peers (Nice80 & Nice60) and then updates the respective learning rates in the RL, of M3. Therefore, the model estimates free parameters separately for mean ( $w_{\text{mean}}, k_{\text{mean}}, \lambda_{\text{mean}}$ ) and nice peers ( $w_{\text{nice}}, k_{\text{nice}}, \lambda_{\text{nice}}$ ).

**M9. Associative Value- Feedback RL Model.** This model assumes that the associability or attention one places on the prediction error vs their already established expectations of the likelihood of receiving nice feedback impacts learning differently based on the valence of the feedback they receive experientially through social interactions, regardless of the social context (i.e. predictability and valence of each peer). Meaning, how much an individual weighs their existing expectations of receiving nice feedback compared to the degree of surprise unexpected nice or mean feedback they just receiving for all peers. To do this, the model estimates two updated associative values (as in M6), one for negative feedback and one for positive feedback trials and then updates the respective learning rates in the RL, of M4. Therefore, the model estimates free parameters separately for mean ( $w_{\text{neg}}, k_{\text{neg}}, \lambda_{\text{neg}}$ ) and nice peers ( $w_{\text{pos}}, k_{\text{pos}}, \lambda_{\text{pos}}$ ).

**M10. Associative Value- Predictability-Valence RL Model.** This model assumes that the associability or attention one places on the prediction error vs their already established expectations of the likelihood of receiving nice feedback impacts learning differently in predictable-vs-unpredictable and mean-vs-nice social contexts. Meaning, how much an individual weighs their existing conceptualization of a peer's reputation compared to the surprise feedback they just received for each peer type. To do this, the model estimates four updated associative values (as in M6), one for predictably

mean peers (Mean80), one for unpredictably mean peers (Mean60), one for predictably nice peers (Nice80), and one for unpredictably nice peers (Nice60) and then updates the respective learning rates in the RL, of M5. Therefore, the model estimates free parameters separately for each peer ( $w_{m8}$ ,  $k_{m8}$ ,  $\lambda_{m8}$ ,  $w_{m6}$ ,  $k_{m6}$ ,  $\lambda_{m6}$ ,  $w_{n8}$ ,  $k_{n8}$ ,  $\lambda_{n8}$ ,  $w_{n6}$ ,  $k_{n6}$ ,  $\lambda_{n6}$ ).

### ***Model Fitting & Comparison***

Parameters were estimated using maximum a posteriori (MAP) estimation to optimize parameters across all choices, using priors of Gamma(1.2, scale = 5) applied to the temperature parameters and Beta (1.1, 1.1) applied to learning rates and the weighting parameter (Hackel, Doll, & Amodio, 2015). To overcome bias of the optimization algorithm to the initial point, the optimization was repeated one-hundred times and the best set of parameters was selected. For each model, the best-fitting parameters were used to compute the Laplace approximation to the Bayesian model evidence (Daw, Delgado, Phelps, & Robbins, 2011). Model evidence was compared using the Bayesian model selection script “bmsR” package in R (Stephan, Penny, Daunizeau, Moran, & Friston, 2009), which uses the Dirichlet parameter, model frequencies, exceedance probabilities (conditional model probability of a model compared to all tested models), Bayesian omnibus risk (i.e. probability that model differences are due to chance), and protected exceedance probabilities (controlling for the null due to chance). The Dirichlet distribution describes the probabilities for all models considered and defines a multinomial distribution over model space, allowing one to compute how likely it is that a specific model generated the data of a randomly chosen subject as well as the exceedance probability of one model being more likely than any other model. Greater

probabilities in each of these measures indicate greater evidence of model fit compared to all other models tested, the null, and chance.

### *Model Parsimony Analyses*

The best model, Model 10, was altered to determine the most parsimonious set of parameters needed for each learning context and then compared to all models to determine the most parsimonious winning model (Wilson & Collins, 2019). In other words, if a different weighting parameter, scaling parameter, and diffusion parameter are necessary for each social context ( $w_{m8}$ ,  $k_{m8}$ ,  $\lambda_{m8}$ ,  $w_{m6}$ ,  $k_{m6}$ ,  $\lambda_{m6}$ ,  $w_{n8}$ ,  $k_{n8}$ ,  $\lambda_{n8}$ ,  $w_{n6}$ ,  $k_{n6}$ ,  $\lambda_{n6}$ ) or if only the weighing parameter should be specific to the reputation contrast (i.e. one  $w$ ,  $k$ , and  $\lambda$  for predictable, or similar valence, or all peers). In line with previous studies of that show learning happens differently in positive-vs-negative feedback contexts (Abler et al., 2006; Jarcho et al., 2015; C. R. G. Jones et al., 2011; R. M. Jones et al., 2011; Maia & Frank, 2011), we also examined if the associability of learning was also specific to positive-vs-negative feedback contexts by using a positive and a negative diffusion parameter for model 10 ( $\lambda_{neg}$ ,  $\lambda_{pos}$ ). Thus, the first parsimonious version of M10, labeled M10  $\lambda_{pos/neg}$ , was identical to the original M10, but instead used a single for scaling parameter ( $k$ ) for all reputations, instead of four reputation-specific scaling parameters ( $k_{m8}$ ,  $k_{m6}$ ,  $k_{n8}$ ,  $k_{n6}$ ), and used two feedback-based diffusion parameters ( $\lambda_{pos}$ ,  $\lambda_{neg}$ ), instead of four reputation-specific diffusion parameters ( $\lambda_{m8}$ ,  $\lambda_{m6}$ ,  $\lambda_{n8}$ ,  $\lambda_{n6}$ ). The second parsimonious version of M10, labeled M10 1W  $\lambda_{pos/neg}$ , was identical to the original M10, but instead used a single for dynamic weighting parameter ( $W$ ) for all reputations, instead of four reputation-specific dynamic weighting parameters

( $W_{m8}$ ,  $W_{m6}$ ,  $W_{n8}$ ,  $W_{n6}$ ), and used two feedback-based diffusion parameters ( $\lambda_{\text{pos}}$ ,  $\lambda_{\text{neg}}$ ), instead of four reputation-specific diffusion parameters ( $\lambda_{m8}$ ,  $\lambda_{m6}$ ,  $\lambda_{n8}$ ,  $\lambda_{n6}$ ).

### ***Simulation: Model-Based Regressor Generation***

To generate parametrically modulated time series for fMRI analysis, the parsimonious winning model (M10  $\lambda_{\text{pos/neg}}$ ) was simulated for each participant using the mean values of the best-fitting parameters ( $W_{m8}$ ,  $W_{m6}$ ,  $W_{n8}$ ,  $W_{n6}$ ,  $k$ ,  $\lambda_{\text{neg}}$ ,  $\lambda_{\text{pos}}$ ) across all participants and participants' individual choice data (true behavioral predictions and actual feedback outcomes). This stabilizes noisy parameter estimates, and the mean of individual parameters provides an estimate of population parameters (Friston et al., 1998). This generated trial-level regressors including 2 types of prediction errors ( $\delta_t$ ; reputation-based and accuracy-based), learning rates ( $\alpha_t$ ), and associative values ( $A_t$ ) for each trial and peer reputation.

Reputation-based prediction-errors ( $\delta_t$ ) were calculated by subtracting the actual feedback received on a trial (1= nice, 0=mean) minus the expected value of the probability of receiving nice feedback for that peer on a trial. Larger positive prediction errors indicate that the feedback received was nicer than the expected from that peer, which in-turn increases the subsequent expected probability of receiving nice feedback from that peer on future trials. Larger negative prediction errors indicate that the feedback received was meaner than the expected from that peer, which in-turn decreases the subsequent expected probability of receiving nice feedback from that peer on future trials. Prediction errors near zero indicate accurate expected probabilities of receiving nice feedback for that peer compared to the underlying true probability of receiving nice feedback from that peer based on their reputation.

Accuracy-based prediction-errors ( $\delta_t$ ) were calculated by subtracting the actual feedback received on a trial (1= nice, 0=mean) minus the expected value of the probability of *being correct about the feedback* for that peer on a trial. Thus, accurate predictions about nice feedback result in identical prediction errors and accurate predictions about mean feedback result in inverse prediction errors for reputation-based and accuracy-based prediction errors.

**Winning Model Validation.** To determine how well the winning model (M10  $\lambda_{\text{pos/neg}}$ ) reproduced actual learning in youth with varying levels of SA we compared learning across each social context over the last 8 trials behavioral with simulated values from M10  $\lambda_{\text{pos/neg}}$ . The last 8 trials were selected because participants should have established reputations for each peer by this time and be making prediction choices in line with the underlying probabilities of receiving nice or mean feedback from a respective peer. To do this, we computed the proportion of nice-vs-mean prediction choices on the last 8 trials for each peer (Mean80, Mean60, Nice80, Nice60), for each participant. Using the best-fitting parameters ( $w_{m8}$ ,  $w_{m6}$ ,  $w_{n8}$ ,  $w_{n6}$ ,  $k$ ,  $\lambda_{\text{neg}}$ ,  $\lambda_{\text{pos}}$ ) *across all* participants for model M10  $\lambda_{\text{pos/neg}}$ , one hundred simulations were done for each participant to produce estimated predictions of each trial. These were then averaged to produce stable individual estimates of predictions on each trial for each participant, which were then used to compute the proportion of nice-vs-mean prediction choices on the last 8 trials for each peer. Average prediction proportions for actual and simulated data were computed for participants with clinically relevant SA ( $\geq 7$ ) vs subclinical or no SA ( $< 7$ ) on the SCARED self-report to compare learning in youth with varying SA and graph within each social context. To validate that the model sufficiently reproduced differences in learning across social contexts and SA, visual inspection of the resulting graph was

used to ensure that simulated values captured the same pattern of prediction choices and fell within the error-bar ranges for each value.

**Learning Analyses.** To understand differences in the mechanisms of social learning across social context factors and varying levels of SA in youth, we used t-tests used to compare winning-model *individual* parameters ( $w_{m8}$ ,  $w_{m6}$ ,  $w_{n8}$ ,  $w_{n6}$ ,  $k$ ,  $\lambda_{neg}$ ,  $\lambda_{pos}$ ) between participants with clinically relevant SA vs subclinical or no SA, using the mean-split of 7 on the SCARED to facilitate interpretations. Next, we assessed how these differences in the mechanisms of social learning in SA produced differences in learning rates ( $\alpha_t$ ), and associative values ( $A_t$ ) across social contexts for those with and without clinical SA symptoms using ANOVAs.

### ***Model-Based fMRI Analyses***

**fMRI Preprocessing.** Data was converted from DICOM to BIDS format using the HeuDicov tool version 0.5.4 and BIDS validator tool version 1.4.0. Standard preprocessing steps were implemented with `afni_proc.py`; these steps included despiking, slice timing, non-linear warping (SS warper), coregistration, smoothing to 6-mm full-width half maximum (FWHM), spatial normalizing to standard space, and resampling, which resulted in 3 mm<sup>3</sup> voxels AFNI software (Cox, 1996).

**Prediction-Error Parametric Modulation.** Prediction-errors were mapped onto trial-level interactions with each peer independently. In situations in which the prediction was nice and the feedback was accurately predicted as nice, the correlation between accuracy-based prediction errors and reputation-based prediction errors would be 1. Since the amount of times this case occurred varies by peer context (ie. Greatest probability for Nice80), modeling all feedback may result in varying levels of highly

correlated prediction error regressors. Indeed, the correlation between the two types of prediction errors was too high across all participants ( $r$ 's  $> 0.402$ ,  $p$ 's  $< 0.001$ ) to consider both the reputation- and accuracy-based prediction errors in the same model. As such, each prediction-error was used in as a modulator of the neural signal during the feedback period in a separate model. Thus, in each model, the feedback onset was parametrically modulated by a time series representing either the reputation- or the accuracy-based prediction error.

The reputation-based prediction-error was coded such that more positive values indicate nicer than expected feedback, more negative values indicate meaner than expected feedback, and values close to zero indicate accurately predicted the probability of receiving nice feedback from a particular peer. The accuracy-based prediction-error was coded such that more positive values indicate feedback in the direction predicted, and negative values indicate feedback in the opposite direction predicted, regardless of the valence of the feedback. Each of these were separately used as linear parametric modulators of the neural signal during feedback in our network of regions of interest. Thus, we are able to interpret increased activation as being linearly correlated with the prediction error on a given trial. Thus, positive correlations indicate increases in brain activation from nicer than expected feedback and negative correlations indicate increases in brain activation from meaner than expected feedback.

**Individual-Level Analyses.** First level analyses were performed using a general linear model (GLM) that included the onset of feedback with a fixed duration of 3 second as the sole regressor of interest, that contain 4 distinct feedback regressors (feedback from: Mean80, Mean60, Nice80, Nice60) and 4 distinct prediction-error parametric modulates that corresponded to each peer regressor. Regressors of no interest included (1) onset of prediction epoch with a duration of the reaction time of the participant's

choice, (2) onset of the response epoch with a duration of the reaction time of the participant's choice, (3 & 4) prediction and response epochs for missed trials (both lasting the duration of the epoch of 4 seconds), (6–11) the six motion parameters (x, y, z, pitch, roll, yaw), (12) linear drift. All variable intra- and inter-trial intervals and the classroom menu epochs were considered part of the implicit baseline. Individual-level regression analyses were carried out with AFNI's 3dDeconvolve function. Temporally adjacent TRs with a euclidean-norm motion derivative  $> 1.0$  mm were omitted from the model via censoring (percent censored;  $M= 4.47\%$ ,  $SD=6.06\%$ ). This resulted in a  $\beta$  coefficient and t statistic for each voxel and regressor. Whole-brain percent signal-change maps were generated by dividing signal intensity at each voxel by the mean voxel intensity, and multiplying by 100.

**Group-level analyses: Valence by Predictability by SA.** To avoid Type I errors in regions where signal dropout occurred, output maps were masked to include voxels where at least 90% of the participants had signal. Analyses were conducted in AFNI's 3dMVM software (Chen, Adleman, Saad, Leibenluft, & Cox, 2014). Two group-level 3-way interaction analyses were conducted, both included two within-subject factors of valence and predictability and one between-subjects factor of SA. The first model, examined the extent to which the reputation-based prediction-error modulated the neural signal in regions of interest varied based on social context (i.e. valence and predictability) and differing levels of SA. The second model, examined the extent to which the accuracy-based prediction-error modulated the neural signal in regions of interest varied based on social context (i.e. valence and predictability) and differing levels of SA. The continuous variable (SA) was centered at 7 to facilitate interpretation, so that values  $\geq 0$  indicate clinically relevant SA and values  $< 0$  indicate subclinical or no SA.

**Apriori Network and Sub-Region Mask(s).** A combined network mask was created comprising of our regions of interested related to learning which included: two bilateral dACC regions, a bilateral vmPFC, bilateral anterior insula, and a bilateral ventral striatum region (Figure 2A-E). The first posterior dACC bilateral 6mm spheres (Figure 2C) were centered around coordinates from our previous work using the Virtual School Task: +/-1, -1, 39 (Jarcho et al., 2016), from which the LEARN task was adapted, in which brain activation related to differences in anticipatory anxiety while anticipating social feedback. The second anterior dACC bilateral 6mm spheres (Figure 2B) were centered around coordinates from a recent meta-analyses examining regions associated with social learning: +/-6, 22, 42 (Feng et al., 2021). Studies vastly differed on coordinates associated with social learning within the vmPFC (Feng et al., 2021). Thus, a highly inclusive mask of the entire vmPFC (Figure 2A) was used to capture all possible clusters modulated by social learning: <https://neurovault.org/collections/5631/> (Bhanji, Smith, & Delgado, 2016). Both social anxiety and social learning have been associated with the anterior insula specifically, rather than the entire insula (Fouragnan et al., 2018; Jarcho et al., 2016; Lindström, Haaker, & Olsson, 2018b; Paulus & Stein, 2006; Rosen et al., 2018; Will, Rutledge, Moutoussis, & Dolan, 2017). Thus, we used a Neurosynth Meta-Analysis of anterior insula (Figure 2D), and a threshold of 10 to mask only the anterior portion of the insula. Lastly, social learning has been isolated to the bilateral ventral striatum (Fouragnan et al., 2018; Garrison, Erdeniz, & Done, 2013b; Lin, Adolphs, & Rangel, 2012; Masten et al., 2009), as such we used the Oxford-GSK-Imanova Striatum Atlas isolating the ventral portion exclusively (Figure 2E).

**Cluster Extraction and Correction.** Given our *a priori* focus on neural circuits related to learning in small regions of interest modulated by learning, we set an initial threshold of  $K_e=10$ ,  $p=.02$  (Roiser et al., 2016) to incorporate activation clusters in small

sub regions of the network mask, such as the bilateral anterior (65 voxels), posterior dACC (46 voxels), and bilateral ventral striatum (107 voxels). We performed cluster corrections using 3dClustSim for the entire network mask (1334 voxels), which resulted in a correction to  $NN=2$ ,  $\alpha=.05$  and  $Ke=40$ . Given that the cluster correction for the entire mask would preclude detecting significant clusters in smaller sub-regions altogether and is mostly comprised of voxels from the bilateral vmPFC sub-region (956/1334 voxels), we also performed cluster correction for each sub-region in line with guidelines for a priori small region of interest analyses (Roiser et al., 2016). This resulted in a correction to  $NN=2$ ,  $\alpha=.05$  in the anterior dACC ( $Ke=7$ ; 65 voxels), posterior dACC ( $Ke=5.5$ ; 46 voxels), bilateral ventral striatum ( $Ke=8.6$ ; 107 voxels), bilateral anterior insula ( $Ke=10.6$ ; 162 voxels), and bilateral vmPFC ( $Ke=39$ ; 956 voxels). We reported all significant clusters with uncorrected p-values as well as denotations of which survived network-based and sub-region-based cluster correction. Significant clusters that emerged from the Valence by Predictability by SA analyses were decomposed for interpretation purposes. Decomposition was performed in RStudio (RStudio Team, 2016) using the “nlme” package for linear mixed effect models and model effects plots were created using the “effects” package for visualizations and interpretation.

## RESULTS

### Model Comparison

Models 1-10 were compared, along with two additional parsimonious versions of the winning model M10. M10  $\lambda_{\text{pos/neg}}$  had the best fit to the data across all metrics (Table 2) and will be used in all further analyses. Based on the parameters of the winning model, M10  $\lambda_{\text{pos/neg}}$  assumes that social learning is influenced by the contextual factors of valence, predictability and the associability placed on their expectation about a peer’s reputation. Specifically, it assumes that learning is dynamically adjusted separately for each reputation of peer (as determined by social context factors of predictability and valence), and that the amount of adjustment of the learning rate for each peer is based on how much surprise a participant is experiencing in either unexpected positive or negative feedback contexts.

*Table 2 Bayesian Model Comparison*

BMS Fit Measure	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M10 1W	
										$\lambda_{\text{pos/neg}}$	$\lambda_{\text{pos/neg}}$	
Dirichlet parameters	1.021	1.013	1.022	10.631	1.405	1.026	1.194	1.117	4.290	13.750	<b>15.246</b>	7.286
Model frequencies	0.017	0.017	0.017	0.180	0.024	0.017	0.020	0.019	0.073	0.233	<b>0.258</b>	0.123
Exceedance probabilities	0.000	0.000	0.000	0.097	0.000	0.000	0.000	0.000	0.001	0.342	<b>0.547</b>	0.013
Bayesian omnibus risk (i.e. probability that model differences are due to chance)	0.000											
Protected exceedance probabilities	0.000	0.000	0.000	0.097	0.000	0.000	0.000	0.000	0.001	0.342	<b>0.547</b>	0.013

***Model Validation of M10  $\lambda_{pos/neg}$***

On average, youth with clinically relevant levels of SA accurately predicted feedback comparable to the true proportions of nice-vs-mean feedback for each peer (Table 3). Whereas, youth with subclinical to no SA tended to over predict nice feedback compared to the true proportions of nice-vs-mean feedback for each peer, aside from the Nice60 where they were very accurate (Table 3). Model simulated prediction proportions were able to capture the same overall pattern in learning and fell within the error bars for each social context (Figure 3), suggesting the model parameters accurately represented the data.

***Table 3. Model Validation: Average Prediction Proportion by Social Context and SA***

<b>Prediction</b>	<b>Mean80</b>		<b>Mean60</b>		<b>Nice80</b>		<b>Nice60</b>	
	<b>Nice</b>	<b>Mean</b>	<b>Nice</b>	<b>Mean</b>	<b>Nice</b>	<b>Mean</b>	<b>Nice</b>	<b>Mean</b>
<b>Low SA</b>	44%	56%	55%	45%	68%	32%	59%	41%
<b>High SA</b>	24%	76%	45%	55%	72%	28%	69%	31%
<b>Actual</b>	20%	80%	40%	60%	80%	20%	60%	40%

***Learning Analyses***

**M10  $\lambda_{pos/neg}$  Parameter Analyses: Mechanism of Learning.** Only the positive feedback-based diffusion parameter ( $\lambda_{pos}$ ) differed by SA ( $p=0.048$ ), such that youth with clinically relevant levels of SA had larger positive feedback-based diffusion parameter estimates. This suggests that youth with clinically relevant levels of SA weigh prediction errors from unexpected positive feedback, regardless of social context (*i.e. for all peers*) less than their existing expectations about receiving nice feedback. Therefore, they update their learning rate *less* from unexpected positive feedback, regardless of the social

context (i.e. valence and predictability of peer reputations). All other model parameters, including learning rates ( $k$ ), dynamic adjustment of learning rates ( $W_s$ ), and the negative feedback-based diffusion parameter ( $\lambda_{\text{neg}}$ ) were comparable across levels of SA ( $p$ 's $>0.151$ ).

**Resulting Differences in Learning Across Social Contexts and SA: Based on Divergent Mechanisms of Social Learning M10  $\lambda_{\text{pos/neg}}$  Parameters.** Model-simulated average learning rates ( $F(1,3)=7.750, p<0.001$ ; Figure 4a) differed by peer reputation and SA. Specifically, youth with clinically relevant compared to non-significant levels of SA had higher learning rates for unpredictably mean, unpredictably nice, and predictably nice peers ( $p$ 's $<0.001$ ), but did not differ in learning rates of predictably mean peers ( $p>.95$ ). Model-simulated average associative values ( $F(1,3)=23.891, p<0.001$ ; Figure 4b) differed by peer reputation and SA. Youth with clinically relevant compared to non-significant levels of SA had higher associative values for predictably nice peers ( $p=0.001$ ), lower associative values for predictably mean peers ( $p=0.001$ ), and did not differ in associative values of unpredictably mean or nice peers ( $p>.21$ ). Higher learning rates translate to greater adjustments of predictions in response to prediction errors and weighting of the magnitude of the prediction error (scaled by the associative value). Lower associative values translate to less adjusting of the learning rates in response to unexpected feedback and more stable learning over time.

### ***Summary of Computational Modeling and Learning Analyses.***

Taken together, these data show similar, yet distinct mechanisms of learning across social contexts and SA in youth. Regardless of SA, the predictability and valence

of the social context as well as the associability placed on their expectation about a peer's reputation all distinctly influence social learning. This finding is evidenced by best-fit model ( $M10 \lambda_{pos/neg}$ ) containing dynamic weighting components for the learning rates for each social context, which did not differ by SA.

Youth with clinically relevant levels of SA accurately learn in predictably mean social contexts and do not adjust their expectations much over time. This finding is evident in the comparable average prediction proportions of nice and mean feedback for predictably mean peers, combined with the small simulated learning rates and associative values to Mean80 peers. Additionally, youth with clinically relevant levels of SA accurately learn in predictably nice, and unpredictably mean and nice social contexts, but do so by rapidly adjusting their learning rates in response to unexpected negative feedback, but not unexpected positive feedback. This is evidenced by the comparable average prediction proportions of nice and mean feedback for predictably nice, and unpredictably mean and nice peers, combined with the larger simulated learning rates and associative values, and larger positive feedback-based diffusion parameter ( $\lambda_{pos}$ ).

In contrast, youth without clinically relevant SA over estimate niceness in all social contexts, expect unpredictably nice contexts. In predictably nice contexts, they do not adjust their expectations much over time. This is evidence by the small simulated learning rates and associative values to Nice80 peers. However, in predictably mean, and unpredictably mean and nice social contexts, they rapidly adjust their learning rates in response to both unexpected positive and negative feedback. This is evidenced by the larger simulated learning rates and associative values for Mean80, Mean60 and Nice60 peers, and smaller positive and negative feedback-based diffusion parameters ( $\lambda_{pos/neg}$ ).

## Model-Based fMRI Analyses

### *Reputation-Based Prediction Error: Valence by Predictability by SA*

Reputation-based prediction-error modulated the neural signal in seven activation clusters within our learning network mask based on the social contexts of valence and predictability across levels of SA (Table 4 & Figure 5A-G). This was primarily driven by a valence by SA interaction within predictable social contexts in vmPFC and right ventral striatum ( $p$ 's < .039; Table 4 & Figure 6A-G). Similar trends emerged in the left anterior insula, left dACC, and left ventral striatum ( $p$ 's < .062; Table 4). Specifically, more severe SA symptoms were associated with *decreased* brain activation to unexpected *negative* feedback in predictably nice social contexts (Nice80; *partial r*'s > -0.299) and unpredictably mean social contexts (Mean60; bilateral vmPFC only; *partial r*'s > -0.067) and *increased* in brain activation to unexpected positive feedback in predictable mean social contexts (Mean80; *partial r*'s > 0.215). No associations were found for unpredictably nice social contexts (Nice60).

To further facilitate interpretation, interactions were also decomposed treating SA as a dichotomous variable, with clinically signifiant and non-significant groups. Results support largely dimensional analyses: youth with compared to without clinically relevant SA had *decreased* brain activation to unexpected *negative* feedback in predictably nice (Nice 80; SA < non-SA; bilateral vmPFC (A), right vmPFC (B), left anterior dACC (C), left ventral striatum (F), and right ventral striatum (G);  $p$ 's < 0.045; Figure 5 & 6), and unpredictably mean (Mean60; SA < non-SA; bilateral vmPFC (A);  $p$  = 0.037; Figure 5 & 6) social contexts.

**Table 4. Group-level Activation Clusters for Reputation-Based Prediction**

**Error: Valence by Predictability by SA.**

Image Key	Network Activation Clusters Modulated by Reputation-based PE	MNI Coordinates			Cluster size (voxels)	Valence*Predictability*SA		Valence*SA	
		x	y	z		F(1,256)	uncorrected p-value	F(1,256)	uncorrected p-value
A	Bilateral vmPFC*+	-10.5	-52.5	-10.5	129.0	9.411	0.002	4.292	0.039
E	Left Anterior Insula*+	31.5	-19.5	-7.5	44.0	7.717	0.005	3.663	0.056
B	Right vmPFC*+	-7.5	-37.5	-22.5	40.0	12.150	<0.001	8.042	0.004
C	Left Anterior dACC+	10.5	-25.5	37.5	21.0	6.273	0.010	3.494	0.062
F	Left Ventral Striatum+	13.5	-13.5	-7.5	15.0	5.478	0.020	3.811	0.052
G	Right Ventral Striatum+	-13.5	-16.5	-7.5	13.0	7.862	0.005	6.834	0.009
D	Right Posterior dACC+	-7.5	1.5	37.5	11.0	5.100	0.024	1.831	0.177

Survives Network-wide 3dClustsim Ke threshold of >40 for alpha=.05 \*

Survives Sub-Region 3dClustsim Ke threshold for alpha=.05 +

**Accuracy-Based Prediction Error: Valence by Predictability by SA.**

The accuracy-based prediction-error did not modulated the neural signal within any of our regions of interest in a 3-way interaction between Valence, Predictability, SA, suggesting that social learning during the LEARN task was driven by the need to understand each peer's reputation, rather than simply being accurate in guessing the type of feedback.

## DISCUSSION

Our results demonstrate that in adolescents, social learning is context dependent and varies based on severity of SA symptoms. This variability manifests behaviorally and is associated with alterations in brain function. Youth with clinically relevant SA learned rapidly and dynamically in predictably nice and unpredictable social contexts by allocating more weight to unexpected negative-vs-positive feedback. This resulted in accurate estimation of peer reputations. Whereas youth without clinical SA learned slowly in all social contexts, and dynamically adjusted learning in unpredictable social contexts by allocating weight to unexpected negative *and* positive feedback. This resulted in a positivity bias in estimating peer reputations. On a neural level, more severe SA symptoms were associated with *decreased* brain activation to unexpected *negative* feedback in predictably nice and unpredictably mean social contexts, and *increased* in brain activation to unexpected *positive* feedback in predictably mean social contexts across salience, reward, and cognitive control regions. Taken together, these data show, for the first time, that fear of social feedback, a core feature of SA disorder, may partially result from differences in how youth learn to interact with others in various social contexts.

### **Mechanisms of Social Learning in Youth**

Social learning in youth relies on reinforcement and associative learning strategies that are context dependent. As evidenced by our best fit model that used distinct dynamic weighting parameters ( $w$ ) for learning rates for each social context (reputation-based valence and predictability), to allow learning rates to adjust differently over time within

each social context. This is consistent with previous findings that social learning is context dependent (Hackel, Mende-Siedlecki, Loken, & Amodio, 2022; Lamba, Frank, & FeldmanHall, 2020), and dynamically adjusts over time through associative learning (Piray et al., 2019). Results build on these findings by demonstrating that context-dependent learning is also a factor in adolescents, is utilized to predict and estimate peer reputations, a skill needed for successful social interactions, and varies based on SA symptom severity.

Replicating previous work highlighting that reinforcement learning occurs differently in positive-vs-negative feedback situations (Maia & Frank, 2011), we found that the best fit model used separate diffusion parameters ( $\lambda_{pos/neg}$ ) for the associability of the prediction error to positive-vs-negative feedback. We extend previous findings by further elucidating the mechanism by which reinforcement learning differs in positive and negative social feedback contexts, through adjusting learning rates over time in relation to the weight allocated to prediction errors from positive and negative social feedback. Despite differences in how learning rates are constructed, the range of values that learning rates could reach did not differ by context or feedback-valence as evidenced by the same scaling parameter ( $k$ ) applied to all learning rates in the winning model. Taken together, this suggests that among youth, learning in social contexts that vary in predictability and valence is a dynamic and context-dependent process that employs associative and reinforcement learning modalities to learn about peers.

### **Distinct Mechanisms of Social Learning Social Anxiety**

The overall mechanisms of social learning in youth were largely similar across levels of SA, however important differences emerged in the weight of unexpected

positive-vs-negative feedback in updating learning rates across social contexts. Specifically, the diffusion parameters for positive feedback ( $\lambda_{pos}$ ) differed by SA, such that in youth with compared to without clinically relevant SA, less weight was placed on unexpected positive feedback in all social contexts when updating learning rates. This is consistent with previous findings suggesting that youth with SA tend to forget unexpected positive feedback (Jarcho et al., 2015; Morrison & Heimberg, 2013), but adds to these findings by elucidating the learning mechanism that facilitates that facilitates this phenomena, and helps explain why youth with SA discount unexpected positive feedback when predicting future social interactions. This critical difference in the role unexpected positive feedback plays in updating dynamic learning over time resulted in clear differences in learning rates, associative values, and accuracy in learning about peers across levels of SA as well as alterations in brain function.

In youth with compared to without clinically relevant SA, learning rates were larger (i.e. faster) in all except predictably mean social contexts; learning rates were small (i.e. slower) for all youth in predictably mean social contexts. This is in part consistent with previous findings in adults with high trait anxiety that show less dynamic adjustments to learning in aversive non-social situations (Browning, Behrens, Jochem, O'Reilly, & Bishop, 2015; Piray et al., 2019), but inconsistent with others that demonstrate less dynamic learning adjustments in unpredictable non-social contexts (Browning et al., 2015). However, our study differed in the social nature of the task, lack of aversive consequence in all unpredictable contexts, was conducted in youth with SA instead of adults with trait anxiety. Thus, social-vs-trait anxiety, as well as age (Smith et al., 2020), may differentially impact learning mechanisms in unpredictable social contexts. More work is needed to tease apart these potential differences.

Associative values differed by level of SA and social contexts. Specifically, youth with clinically relevant SA had large associative values in all except predictably mean social contexts; placing greater weight on unexpected feedback to dynamically update learning, rather than previously established expectations for these contexts. In contrast, youth without clinically relevant SA had large associative values to unpredictable social contexts; but small associative values to predictable social contexts. This suggests that unexpected feedback dynamically updated learning in only unpredictable social contexts. Results are partially consistent with prior studies showing a suboptimal relation between dynamic learning in predictable and unpredictable contexts in adults with high trait anxiety (Browning et al., 2015). However, our findings demonstrate greater adjustments of learning in relation to unexpected feedback in SA, similar to some previous non-social studies (Zika et al., 2022), but in contrast to others that show no or little adjustment of learning in relation to unexpected feedback (Browning et al., 2015; Piray et al., 2019). Thus, youth similarly use prediction errors in unpredictable social contexts to update learning. However, youth with SA uniquely engage this same strategy in predictably nice social contexts, likely due to their emphasis on unexpected negative feedback in learning.

Youth with compared to without clinically relevant SA more accurately estimated the probability of receiving nice feedback from peers in each social context. This is consistent with previous work showing that despite a negativity bias in learning in SA, learning of true probabilities is highly accurate (Zika et al., 2022). Youth with clinically relevant SA achieved accurate learning through discounting unexpected positive feedback and placing greater weight on unexpected negative feedback to adjust learning across social contexts. Whereas youth without clinically relevant SA utilize unexpected positive and negative feedback to adjust learning in unpredictable social contexts. This is consistent with data suggesting healthy individuals have a strong positivity bias in social

learning and individuals with SA tend to lack that positive bias (Button, Browning, Munafò, & Lewis, 2012; Koban et al., 2017; Müller-Pinzler et al., 2019), and extends this work by demonstrating this phenomena, when applied in different social contexts, generates important differences in learning rates and associability of unexpected social feedback.

Taken together, decreased reliance on unexpected positive feedback to adjust learning in SA results in more accurate learning across contexts, but does so by emphasizing recent unexpected negative feedback in predictably nice and unpredictable social contexts. Youth without SA do not learn as accurately, and only rapidly adjust learning in unpredictable contexts in response to unexpected positive and negative social feedback. This distinction in how learning occurs in social situations could result in greater fear-based associations to unpredictable and negative social interactions in youth with SA, which is hallmark to SA disorder.

### **Model-Based Prediction Error Modulated Neural Activation**

Reputation-, but not accuracy-based prediction-errors linearly modulated brain activation in the reward, salience, and shared cognitive control hubs associated with learning in youth. This is consistent with previous work demonstrating linear relations between non-social positive and negative prediction-errors and brain activation in reward and salience circuitry in adults (Behrens et al., 2009, 2008), and extends these findings to social contexts in adolescents. Results also provide further evidence for shared circuitry of learning in social and non-social contexts (Bhanji & Delgado, 2014). At the same time, it suggests that in social situations, the motivation to learn what to expect from peers, rather than the desire to be accurate about their feedback, drives learning. Although this is

somewhat inconsistent with our previous work showing higher reward center activation upon receipt of accurately predicted one-time negative peer feedback in high anxiety youth (Quarmley et al., 2019), the LEARN task prioritized the prediction of future social interactions with the same peers across time, which may have primed learning towards reputation-based prediction errors rather than accuracy. Thus, future studies should clarify when accuracy-vs-reputation contexts drives learning in one-time vs. ongoing social interactions.

### **Neural Mechanisms of Social Learning in Social Anxiety**

Youth differed in the neural mechanisms of social learning across social contexts with varying levels of SA. More severe SA was associated with *decreased* brain activation in predictably nice and unpredictably mean social contexts, and *increased* brain activation in predictably mean social contexts. These findings are somewhat consistent with our previous work demonstrating greater brain activation in salience regions while *anticipating* social feedback in unpredictable social feedback in youth with high risk for SA (Clarkson et al., 2019; Jarcho et al., 2016). However, the present study focused on the feedback period of the social interaction when learning occurs. Thus, our results may provide insight into how *decreased* neural engagement and rapidly increased learning after unexpected negative feedback in unpredictably mean social contexts may potentiate anticipatory anxiety in unpredictable social situations in youth with SA.

Our results are also consistent with our previous work demonstrating increased reward and shared cognitive control center activation to unexpected positive feedback in youth with SA (Jarcho et al., 2015), but further specify this occurs in predictable mean social contexts and is often discounted when using such information to formulate

expectations about future social interactions. Interestingly, SA-related pattern of activation in predictably nice and mean social contexts occurred across salience, reward, and cognitive control regions. However, in unpredictably mean social contexts, decreased activation only occurred in the bilateral vmPFC, a cognitive control region. These findings are consistent with studies in adults that highlight the role of cognitive control over regulating salience and reward response to unpredictable contexts in non-SA individuals (Veit et al., 2012), and evaluating magnitudes of gains-vs-losses (Hollmann et al., 2011). Our results build these findings by demonstrating how, in youth with SA, dysregulation the vmPFC in response to unexpected negative feedback in unpredictably mean social contexts may, in turn, impair regulation of the salience and reward network when anticipating future social interactions with unpredictable peers. This could result in greater anxiety and intolerance of uncertainty, another core feature of SA.

## **Conclusions**

Taken together, our study provides new insight about how youth with clinically relevant SA learn to fear social interactions by emphasizing unexpected negative feedback, and discounting unexpected positive feedback, in predictably nice and unpredictable social contexts to rapidly adjust learning. In contrast, youth without clinically relevant SA use both positive and negative unexpected feedback and adjust learning more in unpredictable vs predictable social contexts. More severe SA was associated with decreased engagement of salience, reward and cognitive control regions in predictably nice social contexts, and only engaged cognitive control regions in unpredictably mean social contexts. Since we measured SA continuously and dysregulation in both neural and behavioral mechanisms linearly related to SA, these

findings could provide evidence for a future susceptibility to social anxiety via learning to fear social interactions in youth.

### **Limitations & Future Directions**

Though our study provides novel and consistent evidence towards dysregulated behavioral and neural mechanisms of social learning in youth with SA, it is not without limitations. Our sample size was small due to complications with collecting data during the COVID-19 pandemic. Thus, results should be interpreted with caution in generalizing to the broader population. However, we used computational modeling to estimate parameters of learning, which allows for greater generalization, even in small samples (Wilson & Collins, 2019), and used highly constrained apriori defined regions of interest to understand neural mechanisms of social learning (Roiser et al., 2016). Our study also modeled feedback across each social context, paring prediction-errors with their respective social contexts. This method confounded accuracy and reputation-based predictions when correctly predicting nice feedback making it difficult to simultaneously test accuracy and reputation-based learning in the same model. Thus, future studies should model feedback within each social context separately using both accuracy and reputation-based modulators to determine unique contributions of accuracy and reputation-derived learning in various social contexts. Finally, longitudinal and/or cross sectional studies should test developmental effects of learning across time to understand SA onset, symptom trajectories, and course, to better understand the causal nature of differences in social learning and SA.

## BIBLIOGRAPHY

- Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *NeuroImage*, *31*(2), 790–795. <https://doi.org/10.1016/j.neuroimage.2006.01.001>
- American Psychiatric Association., A. A., & American Psychiatric Association. DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5*. : American Psychiatric Publishing. <https://doi.org/10.1016/B978-1-4377-2242-0.00016-X>
- Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, *2*(11). Retrieved from <http://neurosci.nature.com>
- Beesdo-Baum, K., Knappe, S., Fehm, L., Höfler, M., Lieb, R., Hofmann, S. G., & Wittchen, H. U. (2012). The natural course of social anxiety disorder among adolescents and young adults. *Acta Psychiatrica Scandinavica*, *126*(6), 411–425. <https://doi.org/10.1111/j.1600-0447.2012.01886.x>
- Beesdo, K., Bittner, A., Pine, D. S., Stein, M. B., Höfler, M., Lieb, R., & Wittchen, H. U. (2007). Incidence of social anxiety disorder and the consistent risk for secondary depression in the first three decades of life. *Archives of General Psychiatry*, *64*(8), 903–912. <https://doi.org/10.1001/archpsyc.64.8.903>
- Behrens, T. E. J., Hunt, L. T., Rushworth, M. F. S., Frith, C. D., Frith, U., Behrens, T. E. J., ... Rushworth, M. F. S. (2009). *The Computation of Social Behavior*. *Science*

- (Vol. 324). <https://doi.org/10.1126/science.1169694>
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. <https://doi.org/10.1038/nature07538>
- Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. <https://doi.org/10.1038/nn1954>
- Bhanji, J. P., & Delgado, M. R. (2014, January). The social brain and reward: Social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*. NIH Public Access. <https://doi.org/10.1002/wcs.1266>
- Bhanji, J. P., Smith, D. V., & Delgado, M. R. (2016). A Brief Anatomical Sketch of Human Ventromedial Prefrontal Cortex. *Nature Neuroscience*, *19*, 1–3.
- Birmaher, B., Brent, D. A., Chiappetta, L., Bridge, J., Monga, S., & Baugher, M. (1999). Psychometric properties of the screen for child anxiety related emotional disorders (SCARED): A replication study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(10), 1230–1236. <https://doi.org/10.1097/00004583-199910000-00011>
- Boehme, R., Lorenz, R. C., Gleich, T., Romund, L., Pelz, P., Golde, S., ... Beck, A. (2017). Reversal learning strategy in adolescence is associated with prefrontal cortex activation. *European Journal of Neuroscience*, *45*(1), 129–137. <https://doi.org/10.1111/ejn.13401>
- Boelen, P. A., & Reijntjes, A. (2009). Intolerance of uncertainty and social anxiety.

- Journal of Anxiety Disorders*, 23(1), 130–135.  
<https://doi.org/10.1016/j.janxdis.2008.04.007>
- Boelen, P. A., Vrinssen, I., & van Tulder, F. (2010). Intolerance of Uncertainty in Adolescents. *The Journal of Nervous and Mental Disease*, 198(3), 194–200.  
<https://doi.org/10.1097/nmd.0b013e3181d143de>
- Boer, J. A. (1997). Social phobia: epidemiology, recognition, and treatment. *BMJ*, 315(7111), 796–800. <https://doi.org/10.1136/bmj.315.7111.796>
- Bolstad, I., Andreassen, O. A., Reckless, G. E., Sigvartsen, N. P., Server, A., & Jensen, J. (2013). Aversive Event Anticipation Affects Connectivity between the Ventral Striatum and the Orbitofrontal Cortex in an fMRI Avoidance Task. *PLoS ONE*, 8(6).  
<https://doi.org/10.1371/journal.pone.0068494>
- Bradford Brown, B., Eichert, S. A., & Petriet, S. (1986). The importance of peer group “crowd” affiliation in adolescence. *Journal of Adolescence*, 9, 73–76. Retrieved from [https://ac-els-cdn-com.libproxy.temple.edu/S014019718680029X/1-s2.0-S014019718680029X-main.pdf?\\_tid=f8b2d5ac-c7e4-4f21-936c-7fac15d3ff72&acdnat=1544220932\\_880926dfcb7b2d1bff463cf8892923be](https://ac-els-cdn-com.libproxy.temple.edu/S014019718680029X/1-s2.0-S014019718680029X-main.pdf?_tid=f8b2d5ac-c7e4-4f21-936c-7fac15d3ff72&acdnat=1544220932_880926dfcb7b2d1bff463cf8892923be)
- Brown, T. A., & Barlow, D. H. (2014). *Anxiety and related disorders interview schedule for DSM-5 (ADIS-5)-adult and lifetime version: Clinician manual*. Oxford University Press.
- Browning, M., Behrens, T. E., Jocham, G., O’Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, 18(4), 590–596.

<https://doi.org/10.1038/nm.3961>

Bruce, S. E., Yonkers, K. a, Otto, M. W., & Eisen, J. L. (2005). Influence of Psychiatric Comorbidity on Recovery and Recurrence in Generalize ... *The American Journal of Psychiatry*, *162*(6), 1179–1187.

<https://doi.org/http://dx.doi.org/10.1176/appi.ajp.162.6.1179>

Button, K. S., Browning, M., Munafò, M. R., & Lewis, G. (2012). Social inference and social anxiety: Evidence of a fear-congruent self-referential learning bias. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(4), 1082–1087.

<https://doi.org/10.1016/J.JBTEP.2012.05.004>

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

Caouette, J. D., Ruiz, S. K., Lee, C. C., Anbari, Z., Schriber, R. A., & Guyer, A. E. (2015). Expectancy bias mediates the link between social anxiety and memory bias for social evaluation. *Cognition and Emotion*, *29*(5), 945–953.

<https://doi.org/10.1080/02699931.2014.960368>

Carleton, R. N., Collimore, K. C., & Asmundson, G. J. G. (2010). “It’s not just the judgements-It’s that I don’t know”: Intolerance of uncertainty as a predictor of social anxiety. *Journal of Anxiety Disorders*, *24*(2), 189–195.

<https://doi.org/10.1016/j.janxdis.2009.10.007>

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., & Cox, R. W. (2014). Applications of multivariate modeling to neuroimaging group analysis: A comprehensive

- alternative to univariate general linear model. *NeuroImage*, 99, 571–588.  
<https://doi.org/10.1016/j.neuroimage.2014.06.027>
- Clarkson, T., Eaton, N. R., Nelson, E. E., Fox, N. A., Leibenluft, E., Pine, D. S., ... Jarcho, J. M. (2019). Early childhood social reticence and neural response to peers in preadolescence predict social anxiety symptoms in midadolescence. *Depression and Anxiety*, (August 2018), da.22910. <https://doi.org/10.1002/da.22910>
- Clarkson, T., Karvay, Y., Quarmley, M., & Jarcho, J. M. (2021). Sex differences in neural mechanisms of social and non-social threat monitoring. *Developmental Cognitive Neuroscience*, 52(November), 101038.  
<https://doi.org/10.1016/j.dcn.2021.101038>
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, 29(3), 162–173. <https://doi.org/10.1006/cbmr.1996.0014>
- Crone, E. A., & Dahl, R. E. (2012). Understanding adolescence as a period of social-affective engagement and goal flexibility. *Nature Reviews Neuroscience*, 13(9), 636–650. <https://doi.org/10.1038/nrn3313>
- Daw, N. D., Delgado, M. R., Phelps, E. A., & Robbins, T. W. (2011). Decision making, affect, and learning: attention and performance XXIII. *Decision Making, Affect, and Learning: Attention and Performance XXIII*, 23, 3–38.
- DeWit, D. J., Chandler-Coutts, M., Offord, D. R., King, G., McDougall, J., Specht, J., & Stewart, S. (2005). Gender differences in the effects of family adversity on the risk of onset of DSM-III-R social phobia. *Journal of Anxiety Disorders*, 19(5), 479–502.

<https://doi.org/10.1016/j.janxdis.2004.04.010>

- Ernst, M., & Fudge, J. L. (2009). A developmental neurobiological model of motivated behavior: Anatomy, connectivity and ontogeny of the triadic nodes. *Neuroscience and Biobehavioral Reviews*. <https://doi.org/10.1016/j.neubiorev.2008.10.009>
- Feng, C., Eickhoff, S. B., Li, T., Wang, L., Becker, B., Camilleri, J. A., ... Luo, Y. (2021). Common brain networks underlying human social interactions: Evidence from large-scale neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *126*(March), 289–303. <https://doi.org/10.1016/j.neubiorev.2021.03.025>
- Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. <https://doi.org/10.1002/hbm.24047>
- Franco Cauda, Andrea E. Cavanna, Federico D'agata, Katiuscia Sacco, SergioDuca, and G. C. G. (2011). Functional Connectivity and Coactivation of the Nucleus Accumbens: A Combined Functional Connectivity and Structure-Based Meta-analysis. *Journal of Cognitive Neuroscience*, *23*(10), 2864–2877. <https://doi.org/10.1162/jocn.2011.21624> T4 - A Combined Functional Connectivity and Structure-Based Meta-analysis M4 - Citavi
- Freitas-Ferrari, M. C., Hallak, J. E. C., Trzesniak, C., Filho, A. S., Machado-de-Sousa, J. P., Chagas, M. H. N., ... Crippa, J. A. S. (2010). Neuroimaging in social anxiety disorder: A systematic review of the literature. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. <https://doi.org/10.1016/j.pnpbp.2010.02.028>

- Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7(1), 30–40. <https://doi.org/10.1006/nimg.1997.0306>
- Garrison, J., Erdeniz, B., & Done, J. (2013a). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*. <https://doi.org/10.1016/j.neubiorev.2013.03.023>
- Garrison, J., Erdeniz, B., & Done, J. (2013b). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*. <https://doi.org/10.1016/j.neubiorev.2013.03.023>
- Groenewegen, H. J., Vermeulen-Van Der Zee, E., Kortschot, A. Te, & Wittex, M. P. (1987). ORGANIZATION OF THE PROJECTIONS FROM THE SUBICULUM TO THE VENTRAL STRIATUM IN THE RAT. A STUDY USING ANTEROGRADE TRANSPORT OF PHASEOLUS VULGARIS LEUCOAGGLUTININ. *Neuroscience* (Vol. 23). Retrieved from [https://ac-els-cdn-com.libproxy.temple.edu/0306452287902752/1-s2.0-0306452287902752-main.pdf?\\_tid=78c3973d-1927-47b5-8630-072caa851822&acdnat=1544498489\\_e440452d26c2732e49ec1d4597eabc2a](https://ac-els-cdn-com.libproxy.temple.edu/0306452287902752/1-s2.0-0306452287902752-main.pdf?_tid=78c3973d-1927-47b5-8630-072caa851822&acdnat=1544498489_e440452d26c2732e49ec1d4597eabc2a)
- Guyer, A. E., Benson, B., Choate, V. R., Bar-Haim, Y., Perez-Edgar, K., Jarcho, J. M., ... Nelson, E. E. (2014). Lasting associations between early-childhood temperament and late-adolescent reward-circuitry response to peer feedback. *Development and Psychopathology*, 26(1), 229–243. <https://doi.org/10.1017/S0954579413000941>
- Guyer, A. E., Choate, V. R., Pine, D. S., & Nelson, E. E. (2012). Neural circuitry

- underlying affective response to peer feedback in adolescence. *Social Cognitive and Affective Neuroscience*, 7(1), 81–92. <https://doi.org/10.1093/scan/nsr043>
- Guyer, A. E., McClure-Tone, E. B., Shiffrin, N. D., Pine, D. S., & Nelson, E. E. (2009). Probing the Neural Correlates of Anticipated Peer Evaluation in Adolescence. *Child Development*, 80(4), 1000–1015. <https://doi.org/10.1111/j.1467-8624.2009.01313.x>
- Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, 18(9), 1233–1235. <https://doi.org/10.1038/nn.4080>
- Hackel, L. M., Mende-Siedlecki, P., Loken, S., & Amodio, D. M. (2022). Context-Dependent Learning in Social Interaction: Trait Impressions Support Flexible Social Choices. *Journal of Personality and Social Psychology*. <https://doi.org/10.1037/pspa0000296.supp>
- Harris, L. T., & Fiske, S. T. (2010). Neural regions that underlie reinforcement learning are also active for social expectancy violations. *Social Neuroscience*, 5(1), 76–91. <https://doi.org/10.1080/17470910903135825>
- Heiser, N. A., Turner, S. M., & Beidel, D. C. (2003). Shyness: Relationship to social phobia and other psychiatric disorders. *Behaviour Research and Therapy*, 41(2), 209–221. [https://doi.org/10.1016/S0005-7967\(02\)00003-7](https://doi.org/10.1016/S0005-7967(02)00003-7)
- Hidalgo, R. B., Barnett, S. D., & Davidson, J. R. T. (2001). *Social anxiety disorder in review : two decades of progress. International Journal of Neuropsychopharmacology* (Vol. 4). Retrieved from <https://academic.oup.com/ijnp/article-abstract/4/3/279/976327>

- Hollmann, M., Rieger, J. W., Baecke, S., Lützkendorf, R., Müller, C., Adolf, D., & Bernarding, J. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE*, *6*(10), e25304. <https://doi.org/10.1371/journal.pone.0025304>
- Ipser, J. C., Stein, D. J., Hawkrigde, S., & Hoppe, L. (2009). Pharmacotherapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews*, (3). <https://doi.org/10.1002/14651858.CD005170.pub2>
- James, A. C., Reardon, T., Soler, A., James, G., & Creswell, C. (2020). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews*, *2020*(11). <https://doi.org/10.1002/14651858.CD013162.pub2>
- Jarcho, J. M., Davis, M. M., Shechner, T., Degnan, K. A., Henderson, H. A., Stoddard, J., ... Nelson, E. E. (2016). Early-Childhood Social Reticence Predicts Brain Function in Preadolescent Youths During Distinct Forms of Peer Evaluation. *Psychological Science*, *27*(6), 821–835. <https://doi.org/10.1177/0956797616638319>
- Jarcho, J. M., Grossman, H. Y., Guyer, A. E., Quarmley, M., Smith, A. R., Fox, N. A., ... Nelson, E. E. (2019). Connecting Childhood Wariness to Adolescent Social Anxiety through the Brain and Peer Experiences. *Journal of Abnormal Child Psychology*, *47*(7), 1153–1164. <https://doi.org/10.1007/s10802-019-00543-4>
- Jarcho, J. M., Romer, A. L., Shechner, T., Galvan, A., Guyer, A. E., Leibenluft, E., ... Nelson, E. E. (2015). Forgetting the best when predicting the worst: Preliminary observations on neural circuit function in adolescent social anxiety. *Developmental*

- Cognitive Neuroscience*, 13, 21–31. <https://doi.org/10.1016/j.dcn.2015.03.002>
- Jefferies, P., & Ungar, M. (2020). Social anxiety in young people: A prevalence study in seven countries. *PLOS ONE*, 15(9), e0239133.  
<https://doi.org/10.1371/JOURNAL.PONE.0239133>
- Jones, C. R. G., Pickles, A., Falcato, M., Marsden, A. J. S., Happé, F., Scott, S. K., ... Charman, T. (2011). A multimodal approach to emotion recognition ability in autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, 52(3), 275–285. <https://doi.org/http://dx.doi.org/10.1111/j.1469-7610.2010.02328.x>
- Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., ... Casey, B. J. (2011). Behavioral and Neural Properties of Social Reinforcement Learning. *Journal of Neuroscience*, 31(37), 13039–13045.  
<https://doi.org/10.1523/JNEUROSCI.2972-11.2011>
- Kessler, R. C. (2003). The impairments caused by social phobia in the general population: implications for intervention. *Acta Psychiatrica Scandinavica*, 108(s417), 19–27. <https://doi.org/10.1034/j.1600-0447.108.s417.2.x>
- Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* *Doi101001archpsyc626617*, 62(6 SRC-GoogleScholar FG-0), 617–627. Retrieved from [http://csaweb105v.csa.com.eres.library.manoa.hawaii.edu/ids70/view\\_record.php?id=18&recnum=31&log=from\\_res&SID=doaebbn8m2ice3fuviou5dv8p7](http://csaweb105v.csa.com.eres.library.manoa.hawaii.edu/ids70/view_record.php?id=18&recnum=31&log=from_res&SID=doaebbn8m2ice3fuviou5dv8p7)

- Klöbl, M., Michenthaler, P., Godbersen, G. M., Robinson, S., Hahn, A., & Lanzenberger, R. (2020). Reinforcement and Punishment Shape the Learning Dynamics in fMRI Neurofeedback. *Frontiers in Human Neuroscience*, *14*, 304.  
<https://doi.org/10.3389/FNHUM.2020.00304/BIBTEX>
- Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., ... Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion (Washington, D.C.)*, *17*(8), 1144.  
<https://doi.org/10.1037/EMO0000296>
- Kuhnen, C. M., & Knutson, B. (2005). The neural basis of financial risk taking. *Neuron*, *47*(5), 763–770. <https://doi.org/10.1016/j.neuron.2005.08.008>
- Lamba, A., Frank, M. J., & FeldmanHall, O. (2020). Anxiety Impedes Adaptive Social Learning Under Uncertainty. *Psychological Science*, *31*(5), 592–603.  
<https://doi.org/10.1177/0956797620910993>
- Lewis, J. D., Evans, A. C., Pruett, J. R., Botteron, K. N., McKinstry, R. C., Zwaigenbaum, L., ... Gu, H. (2017). The Emergence of Network Inefficiencies in Infants With Autism Spectrum Disorder. *Biological Psychiatry*, *82*(3), 176–185.  
<https://doi.org/10.1016/j.biopsych.2017.03.006>
- Li, X., Large, C. H., Ricci, R., Taylor, J. J., Nahas, Z., Bohning, D. E., ... George, M. S. (2011). Using interleaved transcranial magnetic stimulation/functional magnetic resonance imaging (fMRI) and dynamic causal modeling to understand the discrete circuit specific changes of medications: Lamotrigine and valproic acid changes in motor or prefrontal e. *Psychiatry Research - Neuroimaging*, *194*(2), 141–148.

<https://doi.org/10.1016/j.psychresns.2011.04.012>

Lin, A., Adolphs, R., & Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social Cognitive and Affective Neuroscience*, 7(3), 274–281. <https://doi.org/10.1093/scan/nsr006>

Lindström, B., Haaker, J., & Olsson, A. (2018a). A common neural network differentially mediates direct and social fear learning. *NeuroImage*, 167(November), 121–129. <https://doi.org/10.1016/j.neuroimage.2017.11.039>

Lindström, B., Haaker, J., & Olsson, A. (2018b). A common neural network differentially mediates direct and social fear learning. *NeuroImage*, 167, 121–129. <https://doi.org/10.1016/j.neuroimage.2017.11.039>

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*. <https://doi.org/10.1038/nn.2723>

Masten, C. L., Eisenberger, N. I., Borofsky, L. A., Pfeifer, J. H., McNealy, K., Mazziotta, J. C., & Dapretto, M. (2009). Neural correlates of social exclusion during adolescence: Understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience*, 4(2), 143–157. <https://doi.org/10.1093/scan/nsp007>

Monahan, K. C., Guyer, A. E., Silk, J., Fitzwater, T., & Steinberg, L. (2016). Integration of developmental neuroscience and contextual approaches to the study of adolescent psychopathology. *Developmental Psychopathology*, 1–46.

Morrison, A. S., & Heimberg, R. G. (2013). Social Anxiety and Social Anxiety Disorder. *Annual Review of Clinical Psychology*, 9(1), 249–274. <https://doi.org/10.1146/annurev-clinpsy-050212-185631>

- Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keyzers, C., ...  
 Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, 9(1), 1–15. <https://doi.org/10.1038/s41598-019-50821-w>
- Muris, P., Merckelbach, H., Schmidt, H., Mayer, B., & Birgit, M. (1999). The revised version of the screen for child anxiety related emotional disorders (SCARED-R): factor structure in normal children. *Personality and Individual Differences Individual Dif*, 26, 99-112doi10. Retrieved from  
[https://s3.amazonaws.com/academia.edu.documents/42489789/The\\_revised\\_version\\_of\\_the\\_Screen\\_for\\_Ch20160209-6632-11cp30d.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1512616939&Signature=TBkWkj5wgK%2FDgNpdO7ZUAPE5yNg%3D&response-content-disposition=inlin](https://s3.amazonaws.com/academia.edu.documents/42489789/The_revised_version_of_the_Screen_for_Ch20160209-6632-11cp30d.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1512616939&Signature=TBkWkj5wgK%2FDgNpdO7ZUAPE5yNg%3D&response-content-disposition=inlin)
- Murray, E. A., & Wise, S. P. (2010). Interactions between orbital prefrontal cortex and amygdala: Advanced cognition, learned responses and instinctive behaviors. *Current Opinion in Neurobiology*. <https://doi.org/10.1016/j.conb.2010.02.001>
- Nelson, E. E., & Guyer, A. E. (2011). The development of the ventral prefrontal cortex and social flexibility. *Developmental Cognitive Neuroscience*, 1(3), 233–245. <https://doi.org/10.1016/j.dcn.2011.01.002>
- Orr, E. M. J., & Moscovitch, D. A. (2010). Learning to re-appraise the self during video feedback for social anxiety: Does depth of processing matter? *Behaviour Research and Therapy*, 48(8), 728–737. <https://doi.org/10.1016/j.brat.2010.04.004>
- Padilla-Coreano, N., Tye, K. M., & Zelikowsky, M. (2022). Dynamic influences on the

- neural encoding of social valence. *Nature Reviews Neuroscience*. Springer US.  
<https://doi.org/10.1038/s41583-022-00609-1>
- Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226. <https://doi.org/10.1038/nature04676>
- Paulus, M. P., & Stein, M. B. (2006). An Insular View of Anxiety. *Biological Psychiatry*.  
<https://doi.org/10.1016/j.biopsych.2006.03.042>
- Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability: Development of a model for Pavlovian learning. *Quantitative Analyses of Behavior*, *3*, 241–256.
- Piray, P., Ly, V., Roelofs, K., Cools, R., & Toni, I. (2019). Emotionally Aversive Cues Suppress Neural Systems Underlying Optimal Learning in Socially Anxious Individuals. *The Journal of Neuroscience*, *39*(8), 1445–1456.  
<https://doi.org/10.1523/JNEUROSCI.1394-18.2018>
- Poore, J. C., Pfeifer, J. H., Berkman, E. T., Inagaki, T. K., Welborn, B. L., & Lieberman, M. D. (2012). Prediction-error in the context of real social relationships modulates reward system activity. *Frontiers in Human Neuroscience*, *6*, 218.  
<https://doi.org/10.3389/fnhum.2012.00218>
- Quarmley, M. E., Nelson, B. D., Clarkson, T., White, L. K., & Jarcho, J. M. (2019). I Knew You Weren't Going to Like Me! Neural Response to Accurately Predicting Rejection Is Associated With Anxiety and Depression. *Frontiers in Behavioral Neuroscience*, *13*(October), 1–11. <https://doi.org/10.3389/fnbeh.2019.00219>
- Raznahan, A., Lee, Y., Stidd, R., Long, R., Greenstein, D., Clasen, L., ... Giedd, J. N.

- (2010). Longitudinally mapping the influence of sex and androgen signaling on the dynamics of human cortical maturation in adolescence. *Proceedings of the National Academy of Sciences*, *107*(39), 16988–16993.  
<https://doi.org/10.1073/pnas.1006025107>
- Reilly-Harrington, N., & Sachs, G. S. (2006). Psychosocial strategies to improve concordance and adherence in bipolar disorder. *The Journal of Clinical Psychiatry*, *67*(7), 14–19. <https://doi.org/10.4088/JCP.0706e04>
- Rescorla, R. A., & Wagner, A. R. (n.d.). *3 A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*. Retrieved from  
<https://pdfs.semanticscholar.org/afaf/65883ff75cc19926f61f181a687927789ad1.pdf>
- Ressler, K. (2007). Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nature Neuroscience*, *10*(9), 1116–1124. Retrieved from <http://www.nature.com/natureneuroscience/>.
- Roiser, J. P., Linden, D. E., Gorno-Tempinin, M. L., Moran, R. J., Dickerson, B. C., & Grafton, S. T. (2016). Minimum statistical standards for submissions to Neuroimage: Clinical. *NeuroImage: Clinical*.  
<https://doi.org/10.1016/j.nicl.2016.08.002>
- Rosen, M. L., Sheridan, M. A., Sambrook, K. A., Dennison, M. J., Jenness, J. L., Askren, M. K., ... McLaughlin, K. A. (2018). Salience network response to changes in emotional expressions of others is heightened during early adolescence: relevance for social functioning. *Developmental Science*, *21*(3), 1–12.

<https://doi.org/10.1111/desc.12571>

RStudio Team. (2016). RStudio: Integrated Development Environment for R. Boston, MA. Retrieved from <http://www.rstudio.com/>

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience*. <https://doi.org/10.1038/nrn2753>

Smith, A. R., Nelson, E. E., Kircanski, K., Rappaport, B. I., Do, Q. B., Leibenluft, E., ... Jarcho, J. M. (2020). Social anxiety and age are associated with neural response to social evaluation during adolescence. *Developmental Cognitive Neuroscience*, *42*, 100768. <https://doi.org/10.1016/J.DCN.2020.100768>

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. <https://doi.org/10.1016/j.neuroimage.2009.03.025>

Van Den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cerebral Cortex*, *22*(6), 1247–1255. <https://doi.org/10.1093/cercor/bhr198>

Veit, R., Singh, V., Sitaram, R., Caria, A., Rauss, K., & Birbaumer, N. (2012). Using real-time fmri to learn voluntary regulation of the anterior insula in the presence of threat-related stimuli. *Social Cognitive and Affective Neuroscience*, *7*(6), 623–634. <https://doi.org/10.1093/scan/nsr061>

Will, G. J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *ELife*, *6*, 1–21.

<https://doi.org/10.7554/eLife.28098>

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *ELife*, 8. <https://doi.org/10.7554/ELIFE.49547>

Yang, X., Liu, J., Meng, Y., Xia, M., Cui, Z., Wu, X., ... He, Y. (2017). Network analysis reveals disrupted functional brain circuitry in drug-naive social anxiety disorder. *NeuroImage*. <https://doi.org/10.1016/j.neuroimage.2017.12.011>

Yin, H. H., Mulcare, S. P., Hilário, M. R. F., Clouse, E., Holloway, T., Davis, M. I., ... Costa, R. M. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience*, 12(3), 333–341. <https://doi.org/10.1038/nn.2261>

Zika, O., Wiech, K., Reinecke, A., Browning, M., Schuck, N. W., & De, Z.-B. M. (2022). Trait anxiety is associated with hidden state inference during aversive reversal learning. *BioRxiv*, 5700, 2022.04.01.483303. <https://doi.org/10.1101/2022.04.01.483303>