A generalizable framework for measuring emotion's role in social decision-making

by

Joseph Heffner

B. Sc., University of North Carolina at Chapel Hill, 2013

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Cognitive, Linguistic & Psychological Sciences at Brown University

Providence, Rhode Island

July 2022

Joseph Heffner

Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Providence, RI 029012 joseph_heffner@brown.edu

EDUCATION

PhD in Psychology, Brown University Expected 2022
Supervisor: Dr. Oriel FeldmanHall

BS in Psychology (with Honors), BA in Philosophy, University of Chapel Hill at North Carolina 2009 – 2013
Supervisor: Dr. Keith Payne; Thesis: "Relevance of emotions in morality: Emotion differentiation moderates influences of disgust on moral judgments"

PUBLICATIONS

1. **Heffner, J.,** & FeldmanHall, O. (revise and resubmit). Model-based planning tunes cooperation rates depending on the context. *Journal of Experimental Psychology: General*.
2. **Heffner, J.**, & FeldmanHall, O. (2022). A probabilistic map of emotional experiences during competitive social interactions. *Nature Communications,* 13(1), 1-11
3. **Heffner, J.,** Son, J., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behavior. *Nature Human Behavior*, 5, 1391-1401
4. **Heffner, J.,** & Vives, M., & FeldmanHall, O. (2021). Anxiety, gender, and social media consumption predict COVID-19 emotional distress. *Humanities and Social Science Communications, 8*, 140
5. **Heffner, J.,** & Vives, M., & FeldmanHall, O. (2021). Emotional responses to prosocial messages increase willingness to self-isolate during the COVID-19 pandemic. *Personality and Individual Differences, 170*
6. **Heffner, J.,** & FeldmanHall, O. (2019). Why we don't always punish: Preference for non-punitive responses to moral violations. *Scientific Reports*, *9*, 13219
7. FeldmanHall, O., Son, J., & **Heffner, J.** (2018). Norms and the flexibility of moral action. *Personality Neuroscience, 1,* E15.
8. Schaich Borg, J., Srivastava, S., Lin, L., **Heffner, J.,** Dunson, D., Dzirasa, K., & Lecesa, L. (2017). Rat intersubjective decisions are encoded by frequency-specific oscillatory contexts. *Brain and Behavior,* 7(6), 1-25
9. Strohminger, N., Gray, K., Chituc, V., **Heffner, J.,** & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Method*s, 48(3), 1187-1204

INVITED TALKS

1. **Heffner, J.** (2022, April). *Economic inequality corrupts moral standards*. Talk presented at Research Matters, Providence, RI.
2. **Heffner, J.,** & FeldmanHall, O. (2022, March). *Modeling emotions using computational tools and machine learning*. Talk presented at the Society for Affective Science Conference, online.
3. **Heffner, J.** (2021, April). *Reverse engineering emotions during social interactions*. Talk presented at the Society for Affective Science Conference, online.
4. **Heffner, J.** (2020, December)*. Emotion prediction errors guide socially adaptive behavior*. Talk presented at Social and Cognitive Brown Bag Series at Brown University, Providence, RI.
5. **Heffner, J.**, (2020, October). *Emotion prediction errors guide socially adaptive behavior.* Talk presented at the Society for Neuroeconomics, online.
6. **Heffner, J.** (2020, November). *Emotion prediction errors guide socially adaptive behavior.* Talk presented at Harvard University, Cambridge, MA.
7. **Heffner, J.** (2020, July). *Emotional mechanisms support prosocial, but not threat, appeals regarding COVID-19.* Talk presented at the SocialBRIDGES Conference, online
8. **Heffner, J.** (2020, June). *Emotion prediction errors guide socially adaptive behavior*. Talk presented at Brown Unconference, Providence, RI.
9. **Heffner, J.** (2019, September)*. Development of wealth-based obligations*. Talk presented at Developmental Brown Bag Series at Brown University, Providence, RI.
10. **Heffner, J.** (2019, July). *The power of emotion prediction errors (and not reward) in biasing social choice.* Talk presented at Northeastern University, Boston, MA.

11. **Heffner, J.** (2018, June). *Emotion prediction errors motivate decisions to punish*. Talk presented at the New England Research on Decision Making, Cambridge, MA.
12. **Heffner, J.** (2017, March). *The power of turning the other cheek: The utility of non-punitive responses to fairness violations*. Talk presented at the New England Research on Decision Making, Providence, RI.

POSTER PRESENTATIONS

1. **Heffner, J.,** Vives, M., & FeldmanHall, O. (2022, April). *Variation in emotion representations predict prosocial choice.* Poster presented at the Society for Affective Science Conference, online.
2. **Heffner, J.,** Son, J., Fan, H., & FeldmanHall, O. (2019, May). *The joy of punishment: Early emotional reactions predict decisions to restore justice.* Poster presented at the Social and Affective Neuroscience Society, Miami. FL.
3. **Heffner, J.,** & FeldmanHall, O. (2018, April). *Emotion prediction errors predict decisions to punish*. Poster presented at the Society for Affective Science, Los Angeles, CA.
4. **Heffner, J.,** Son, J., & FeldmanHall, O. (2018, March). *Emotion prediction errors predict decisions to punish.* Poster presented at the Mind Brain Research Day, Providence, RI.
5. **Heffner, J.,** FeldmanHall, O. (2018, March). *The Power of Turning the Other Cheek: The Limited Value of Punishment*. Poster presented at the Society for Personality and Social Psychology, Atlanta, GA.
6. **Heffner, J.,** & FeldmanHall, O. (2017, March). *Turning the other cheek: The limited utility of punishment*. Poster presented at the Mind Brain Research Day, Providence, RI.
7. **Heffner, J.,** Cameron, D., & Payne, K. (2013, April). *Relevance of emotions in morality: Emotion differentiation moderates influences of disgust on moral judgments*. Poster presented at the Celebration of Undergraduate Research, University of North Carolina at Chapel Hill, NC.

GRANTS AND AWARDS

Graduate Award in Brain Science at Brown University ($57,000) 2020-2021

Carney Institute Brain Science Reproducible Paper Prize at Brown University ($250) 2020

Travel Award SPSP Annual Convention ($500) 2018

Duke Institute for Brain Science (DIBS) Incubator Award ($75,000) 2015 – 2016
Measure Implicit Moral Attitudes Renewal; Cowrote with Dr. Walter Sinnott-Armstrong

Duke Institute for Brain Science (DIBS) Incubator Award ($120,000) 2014 – 2015
Measure Implicit Moral Attitudes; Cowrote with Dr. Walter Sinnott-Armstrong

TEACHING EXPERIENCE

Teaching Assistant/Guest Lecturer, Making Decisions 2022
Brown University, Cognitive, Linguistics, and Psychological Sciences Department

Teaching Assistant/Guest Lecturer, Research Methods 2021
Brown University, Cognitive, Linguistics, and Psychological Sciences Department

Teaching Assistant/Guest Lecturer, Core Concepts in Cognitive and Psychological Sciences II 2018
Brown University, Cognitive, Linguistics, and Psychological Sciences Department

Teaching Assistant/Guest Lecturer, Human Cognition 2018
Brown University, Cognitive, Linguistics, and Psychological Sciences Department

Teaching Assistant/Guest Lecturer, Making Decisions 2017
Brown University, Cognitive, Linguistics, and Psychological Sciences Department

Teaching Assistant, Psychological and Philosophical Summer Studies — 2011 – 2013
Duke University, Duke Talent Identification Program

DEPARTMENTAL SERVICE

Member of Diversity and Inclusion Action Plan, Brown University — 2020 – 2022
Colloquium Organizer, Brown University — 2019 – 2020
Social Cognitive Science Seminar Organizer, Brown University — 2018 – 2019
Graduate Student Council (GSC) Representative, Brown University — 2018 – 2019

AD HOC REVIEW (* indicates co-reviewer with mentor)

Applied Cognitive Psychology
Biological Psychology*
Current Directions in Psychological Science*
Cyberpsychology
Evolution and Human Behavior
IJHPM
Journal of Experimental Psychology: General
Journal of Experimental Social Psychology
Judgment and Decision Making*
Nature Human Behavior*
Neuropsychopharmacology*
PNAS*
Psychological Review
Psychological Science*
Scientific Reports

MENTORING

2016 – 2022 **Brown University.** Providence, RI.

Grace Denning – Honors thesis, Winner of Muriel Fain Sher Premium for Excellence in Psychology for this work
Willy Lee – Honors thesis
Skye Harris – Independent research study affiliated with Princeton Day High School
Eleanor Gorham – Independent research study affiliated with School Without Walls High School
Deeksha Cirigadi – Action Potential Advising Program high school mentee
Gregory Wu – Action Potential Advising Program high school mentee

Acknowledgements

Oriel, I am extremely grateful for all the time and effort you have spent helping me become a better scientist. You taught me how to ask deep research questions, push through difficult challenges that arise, and importantly, the art of crafting a compelling narrative. You are an inspiration for early career researchers of how to blaze your own path in science, and I am incredibly lucky to have experienced that first hand. Thank you for everything, and I look forward to the future research we do together.

Thank you to my defense committee, Amitai and Michael, who generously provided feedback at multiple stages of this work. You both were patient in answering my questions, even about the smallest details, and taught me how to evaluate emotion concepts and build computational models. It takes a department to raise a graduate student, and you both have helped shape me into the researcher I am today.

To the FeldmanHall Lab, thank you for all the countless moments of joy and laughter. Graduate school can be taxing, but I have always felt like the lab was a second home. To Jae, thanks for the late night conversations over Wong's. To Amrita, thanks for being my desk neighbor and helping me with computational models. To Marc, thanks for keeping science fun and exciting. To Logan, thanks for the early morning talks. To Carolyn, thanks for being there for me whenever I needed an afternoon walk. To Harrison, thanks for pushing me to refine and improve my ideas. To James, for my support pillar and coffee buddy. Thanks to the CLPS department and my cohort, especially Meghan, Emily, Joey, Trent, Danny, Alana, Dan, Pete, Jeroen, Steve, David, Romy, and Andrew for the good memories and helping me out along the way. Lastly, to the SOMA lab and Frederike, thank you for allowing me to be a part of your amazing science group.

To my mother, thank you for encouraging my love of reading and discussion of social issues. To my father, thank you for teaching me mathematics, logic, and how to have a kind heart. To Ryan, you showed me how to forge your own path and I'm grateful to always have you to turn to.

Lauren, you have been a rock for me during these years and I could not have done this without your unending support. Thank you for keeping me grounded and helping me make the most of our life together.

Contents

List of Tables

List of Figures

Abstract of "A generalizable framework for measuring emotion's role in social decision-making" by Joseph Heffner, Ph.D., Brown University, July 2022.

Emotions are ubiquitous in all human experiences, affecting every part of our social lives including our thoughts, perceptions, and decisions. When faced with someone asking you for money on the street, you might feel overwhelming sadness or reflect on the feeling of guilt that may come if you choose to ignore the request. The immediate emotions (sadness) and expected emotions (guilt) are two possible ways emotions enter the decision-making process and shape future behavior. Theories of decision-making largely focus on the role of discrete emotions motivating choice, such as how anger motivates punishment or sadness increases altruism. One challenge for measuring specific emotions is that emotions are rarely homogenous, and any account of emotion's role in social choice must be able to draw general conclusions beyond specific emotion states. Across four chapters, this thesis provides an empirical and generalizable account of emotion's role in decision-making, by combining a low-dimensional, descriptive map of core affect (i.e., emotion variability on dimensions of unpleasantness and intensity) with popular social choice paradigms. Using machine learning algorithms and a data-driven approach, I show that a diverse set of negative emotions, not only anger, motivates a host of punitive decisions. Extending this framework to real-world decisions about COVID-19, I find that prosocial and threat-based interventions achieve compliance through different emotional mechanisms. I next develop a novel framework for mathematically computing violations of emotion expectations—emotion prediction errors (PEs)—and test how emotion PEs, compared to reward PEs, drive social decision-making. Results demonstrate that emotion PEs have independent, and stronger, contributions to choice than reward PEs and that individuals at risk have selective impairments in the use of emotion, but not reward, PEs. Finally, I use EEG to investigate the neural correlates of emotion and reward PE processing. I find that there is a dissociable neural encoding of reward PEs, indexed by the FRN, and emotion PEs, indexed by both subcomponents of the P300. Together, this work provides an empirical account for how immediate and expected emotions affect choice, and offers a flexible yet unified framework to examine the role of emotion in decision-making.

Chapter 1

General Introduction

Emotions and social decision-making

In 2014 Jon Lester was the starting pitcher for the Boston Red Sox and an All-Star Game player who had just pitched his team to a World Series title the year prior. Lester had been an integral part of the Red Sox winning multiple championships, and now his current contract was coming to an end. In negotiating a new contract, the Red Sox offered Lester a paltry $70 million, less than half the package of other comparable pitchers. Lester was so offended that he shut down subsequent negotiations and decided to go on the open market, where he subsequently signed a deal with the Chicago Cubs for more than twice that, cashing in on a $155 million deal (Shaughnessy, 2014). Although Lester expected a reasonable contract from his beloved hometown team, his surprise and emotional reaction to the low-ball offer changed his priorities and led him to cut all ties with the franchise he had been loyal to for more than a decade. What aspect of Lester's emotional experience altered the decision-making process?

Social interactions are the most likely to elicit strong emotional responses in ourselves as we decide how to respond to an embarrassing situation, confess feelings to a romantic partner, or tell a friend an uncomfortable truth. Despite this, relatively few studies have examined the role of emotion during social decision-making and learning. This is partly because the social world is complex and involves accounting for constructs not always seen in non-social decisions such as mental state inferences (Lee & Harris, 2013) or alternative goals (Deci & Ryan, 2000). In addition, the lack of interplay between affective science that focuses on mechanistic models of emotion, and reinforcement learning frameworks that largely treats emotions as nuisance variables in the choice process leaves gaps in our understanding of the role of emotions in decision-making. In my dissertation, I address this gap by developing a generalizable framework for emotion that can enable us to more deeply understand how subjective feelings impact social choices.

Theories of emotion

"I know it when I see it"
Justice Potter Stewart, *Jacobellis v. Ohio*, 378 U.S. 184 (1964)

The paradoxical nature of emotions is that on the one hand, everyone has an intuitive understanding of emotions (i.e., folk theories of emotions; Johnson-laird & Oatley, 1992), but after over a century of theorizing and scientific debate, there is little consensus about how to precisely define emotions (Barrett et al., 2016). Within this rich history there are three main eras of the scientific examination of emotions: the "golden years" (1855-1899), the "dark ages" (1900-1959), and the "renaissance" (1960s - present; Gendron & Feldman Barrett, 2009) .

Basic emotion theories

The golden years began to lay the framework for the "basic emotion" approach in understanding certain emotions as special kinds of biologically driven responses triggered by objects and events in the world (Tomkins, 1962). The adjective basic is meant to communicate two important features of the theory: 1) there are a set of separate emotions which differ from each other in important ways, and 2) these emotions evolved to serve an adaptive, functional role during evolution in shaping the characteristics of these emotions (Ekman, 1992b). While the number of basic emotions seems to be growing (e.g., Cowen & Keltner, 2017), most basic emotion theorists recognize six emotions that serve as the core foundation which include anger, fear, happiness, sadness, disgust, and surprise (Ekman & Friesen, 1971). These six emotions are distinguished from other affective phenomenon because each is thought to have unique features regarding its visible signal, physiology, and antecedent events. For example, the Facial Action Coding System, a popular tool that breaks facial expressions into individual components of muscle movements called action units, classifies anger based on the unique combination of lowering the brow, raising and tightening the upper lip (Ekman & Friesen, 1978). Researchers have documented the uniqueness of the basic emotions by showing that people across cultures can reliably classify these emotions from photographs of prototypical emotional facial expressions (Matsumoto et al., 2008).

Emotions such as anger and happiness have a privileged status in this tradition, as they are thought to be the products of evolutionary mechanisms, sometimes called affect programs (Ekman & Cordaro, 2011), which are biologically distinct, preprogrammed, and involuntary, and represent both the cause of the emotional experience and the criterion for identifying emotional episodes (Zachar, 2021). Given this logic, basic emotion theorists search for the evolutionary significance of basic emotions and how they might be related to the ability to survive. For example, it is contended that happiness is needed for reproducing, fear is required for the fight-or-flight response, anger serves to assert social dominance, and so on (Plutchik, 1962). Other emotions not included in this primitive set (e.g., rage, frustration) are thought to exist under the hood of one the basic emotions (e.g., anger), and are sometimes called secondary emotions (Ekman et al., 1987). Researchers vary in the extent that they consider emotions other than the basic six and often assume that other emotion states are simply combinations of the basic six (Du et al., 2014; Ortony & Turner, 1990).

One of the strongest challenges against the basic emotion theory is that emotional responses are rarely homogenous (i.e., anger can feel and look very different, both within a person and between people). Basic theorists try to address this by identifying a common theme for a family of emotions (typically shaped by evolution). For example, in early work identifying muscle movements of anger, face researchers identified more than 60 expressions of anger (Ekman & Friesen, 1975), but argued they shared common set of muscular patterns (e.g., lowered brow). In recent work, identifying a common suite of emotions across diverse experiences often involves positing additional factors that might influence a person's experiences such as cultural norms (Ekman, 1992a; Keltner et al., 2019). For example, some researchers propose that variations in emotion expression may be thought of like dialects (Elfenbein, 2013), where deviations from a prototypical emotion (or language) might be accounted for by cultural differences (Cordaro et al., 2018).

After years of basic emotion research, the field fell into the Dark Ages when relatively little was published on emotions. The Dark Ages is largely attributed to the rise of behaviorism, a theory which bypasses intermediate feelings or states of the mind by focusing directly on the physical causes of behavior (Skinner, 1974). However, variability in physical observations of emotions in the body, face, and behavior was documented, and researchers began questioning the founding tenants of basic emotion research. The inability to find discriminable responses for each basic emotion category meant that few models of emotion were fully developed during this time (Gendron & Feldman Barrett, 2009), critiques that are often reflected in later emotion theories developed during the emotion Renaissance.

Appraisal theories

Emotion research has exploded in popularity since the 1960s, driven by work focusing on the appraisal of emotion (Arnold, 1960) and the two-factor theory of emotion (Schachter & Singer, 1962). The basic premise of appraisal theory is that emotions are the result of various cognitive appraisals of stimuli which gives them meaning in relation to the observer (Arnold, 1960; Ellsworth, 2013). In other words, emotions arise from a person's perception of the environment and how the emotion relates to their well-being. While this has the potential to be vague, one core focus of appraisal theorists is to identify the continuous elements which go into the construction of meaning – the appraisals. Debate centers around the specific appraisal factors incorporated under each theory, but some examples include goal relevance (i.e., "how relevant is this situation to my needs?"), degree of control and agency (i.e., "how much control/agency do I have in this situation?"), and unexpectedness (Moors, 2017).

To illustrate the appraisal perspective, imagine a person is aimlessly walking down campus when a squirrel jumps out in front of them causing them to focus their attention on the novel stimulus. The appraisal of novelty is associated with physiological changes such as widening of the eyes or other motor changes (Ellsworth, 2013). In addition, the appraisal of valence may be simultaneous with the appraisal of novelty (Zajonc, 1980), such that the same squirrel can evoke different emotional experiences depending on the individual. If this particular person really likes squirrels, then they might start to experience interest or attention initially, which develops into joy or curiosity or other emotions consistent with positive ongoing appraisals. Other individuals may deem the novel stimulus irrelevant, and thus other relevant, emerging emotional experiences might stop. Appraisal theorists view this as a recursive and continual process, as all components of the emotional experience including appraisals, subjective feelings, physiological responses, and others exert bidirectional influence on each other.

Constructivism theories

The other major theory of emotion that started in the 1960s, born out of Schachter & Singer's classic experiment, was the beginning of the constructivism view of emotion. In their experiment, participants were given epinephrine (adrenaline), a hormone that increases arousal by increasing heart rate and blood pressure, and were either informed about these side-effects or not given any information. Those who were informed had a cognitive explanation for their arousal state, while the uninformed participants' unexplained arousal led them to be more susceptible to interpreting these bodily feelings as an emotion stemming from their own experiences (i.e., either anger or euphoria, depending on how others around them behaved). This experiment powerfully demonstrated the interaction between physiological arousal and cognitive interpretations of that arousal. Over time, this idea was further refined and postulates two stages in generating an emotion: 1) a consciously accessible neurophysiological state characterized by two dimensions: arousal (intensity) and valence (pleasantness/unpleasantness), as well as 2) social construction (Russell, 2003). This two-dimensional neurophysiological state is thought to underly any emotional experience and is referred to as core affect. Changes in core affect can result from internal or external events, and unlike the basic emotion theories, represents a continuous assessment of one's affective state along the dimensions arousal and valence. Social construction refers to the idea that once core affect is attributed to a cause, an emotional episode is constructed in relation to the person's social customs and cultural ideas, representing an instantiation of a categorical emotion such as anger or fear.

Both appraisal and constructionist models highlight cognitive interpretations as a key element to creating meaning during emotional experiences (Oatley & Johnson-Laird, 2014). Because of this, some researchers believe that the appraisal and constructionist views are compatible (Ortony & Clore, 2015), and that they merely focus on different non-emotional constituents of the emotional process. However, if there was a potential distinguishing feature it might be the emphasis on internal versus external factors of an emotional episode. Appraisal models emphasize the meaning of the situation in relation to the person, and the internal physiological changes follow. In psychological constructionism, the emphasis is on making internal or affective states meaningful, such that the first element of an emotional episode is the core affect state. There are many ways to carve the world at its joints, and in general these theories share more commonalities than differences.

Measuring emotion

The debate surrounding emotion theories is also observed when researchers discuss how to measure emotions. Most theories recognize that emotions are individual and subjective, and as such, are both difficult to define and measure. Researchers have measured emotion by asking about subjective feelings, recording physiological changes, and documenting how the face and body expresses itself (Frijda, 1986). The goal of emotion measurement is to capture variations in magnitude of the emotional experience (Plutchik, 1989), but this often requires focusing on a specific aspect of emotion, which has led theorists to prioritize certain measures over others. In this section, I outline popular techniques that have been used to advance emotion theories and evaluate their costs and benefits.

Basic emotion measurements

Basic emotion theorists use discrete or categorical emotion ratings to capture specific emotions including the six basic ones: anger, surprise, disgust, happiness, fear, and sadness (Ekman, 1989). Mood checklists are a common way of measuring emotional experiences and participants are asked to select one (or more) adjectives (e.g., *calm, joyful, disappointed*) which best describe their immediate, subjective feelings. These direct and straightforward measures have the advantage of capturing a person's immediate self-reported emotions, and have been used to uncover the emotions experienced in everyday life (Trampe et al., 2015) and even the structure of emotion concepts (Cowen & Keltner, 2017).

Although discrete emotion terms, like anger, are important for the communication and description of one's subjective feeling states, thousands of semantic terms exist for describing how people feel. Self-report checklists often contain only a fraction of these (e.g., PANAS; Watson et al., 1988), and these measures are specialized to researchers' specific emotion interests. For example, researchers interested in mapping emotions in everyday life include 18 distinct emotions that are measured over approximately 35 days (Trampe et al., 2015). Six of these emotions are only experienced at or less than 5% of the time (*contempt, embarrassment, offense, awe, fear, guilt*). While the results still demonstrate the frequency of those specific emotions, it highlights a major constraint: selecting a set of emotion terms limits the generalizability of results.

People use emotion terms to categorize their state (e.g., I'm feeling happy vs sad), but it's often recognized that these experiences vary in degree and can exist on a continuum (e.g., happy – sad continuum). While there is some debate about how to measure emotions on a continuum (Stevens, 1946), modern researchers largely agree that all, or almost all, psychological constructs can be measured on an ordinal (Likert) scale. To account for variation in magnitude of emotional experience, some measures add quantitative elements by measuring the degree of experiencing a discrete emotion. For example, the positive and negative affect schedule (PANAS; Watson et al., 1988) measures variation in general experience of carefully selected emotional states (e.g., *distressed, scared, hostile, proud, excited*) on a feeling continuum of not at all experienced (1) to extremely experienced (5). Researchers average across all emotions to arrive at composite scores for positive and negative activation, which mitigates the lack of generalizability when using categorical measures. However, even when the emotional context elicits these discrete valenced emotions, the composite measures may not only track positive and negative affect. For example, anger, a prototypical negative emotion, positively correlates with the composite positive arousal score from PANAS, suggesting that this subscale might tap non-valence approach motivation (Harmon-Jones et al., 2009). While discrete emotion measures offer a useful tool for studying specific emotions, they are difficult to compare across contexts and often challenge the generalizability of the results.

Appraisal and constructivist measures of emotion

Constructivists and appraisal models of emotions tend to favor dimensional measures, where the goal is to represent the underlying dimension of the emotion space. As mentioned above, decades of empirical work demonstrate that two dimensions consistently account for a large proportion of the variance when comparing similarity between emotion terms: 1) valence: where states are characterized by a continuum of pleasant to unpleasant feelings, and 2) arousal: a continuum of low to high energy or intensity of the emotional state. Together, these dimensions constitute *core affect* and represent a prerequisite for emotional experiences (Russell, 2003). An important clarification for these theories is that core affect is a necessary condition (meaning required), but not sufficient condition, for emotional experiences. In other words, all emotional experiences involve core affect but core affect on its own is not necessarily enough to constitute an emotional experience. Changes in core affect that are not attributed to situations or contextualized are experienced as mood but if core affect is attributed to a specific event, then it is classified as an emotional experience.

To give an illustrative example of how core affect is involved in an emotional experience let's consider watching a scene from a terrifying movie. The scene (antecedent event) is perceived in terms of its affective qualities (e.g., the low lighting, tense music, deliberate pacing) and changes in core affect occur continuously as the scene unfolds (e.g., experiencing more intensity of unpleasant feelings). Core affect changes as a direct result of the unfolding scene, yet behavior is not altered by the film as people stay glued to their seats waiting to see the scene resolve. Some philosophers call these feelings "proto emotions" (Robinson, 1995), but to constructivists, the experience of an emotion such as fear must be related to the attribution of core affect to the emotional scene. While the fast heartrate, sweaty palms, and rapid breathing related to watching the scene is also found in fear experiences, these alone are not sufficient to be labeled an emotional episode of fear. Emotional states require core affect plus some other non-emotional cognitive process, such as conceptual knowledge about emotions in context (Lindquist, 2013; Smith & Kirby, 2001).

The way that people conceptually represent emotions along two primary dimensions of valence and arousal plays a large role in how they experience the emotions they feel (Satpute et al., 2016).

This is evident in early development as children as young as three typically categorize facial expressions as either positive or negative (i.e., they attend only to valence), and gradually learn to parse emotions into more specific categories as verbal knowledge develops (Nook et al., 2017; Widen, 2013). By age eight, children report simultaneously feeling positive and negative emotions (Larsen et al., 2007). People's ability to identify and differentiate between emotions is documented in people with depression (Demiralp et al., 2012), social anxiety disorder (Kashdan & Farmer, 2014), and individuals who engage in binge drinking and self-injurious behavior (Kashdan et al., 2015). These results demonstrate how changes in core affect impact the ability to differentiate between emotions, and are implicated in mental health disorders.

When comparing discrete or categorical and dimensional methods, one important consideration is the degrees of freedom available in each approach. In a recent study, researchers collected self-reports of emotional states elicited by 2,185 short emotional videos and measured both 31 discrete emotions and 14 dimensions of emotional experience (Cowen & Keltner, 2017). The dimensional approach had participants rate their feelings on various cognitive appraisals including feelings of safety, control, valence, arousal, and others. Researchers used both of these measures to investigate how emotions relate to each other in an abstract representational state space that captures variation in emotion-related responses (Cowen & Keltner, 2021). They then used statistical techniques to quantify this variability and concluded that there were 27 distinct varieties of emotional experience (stemming from the discrete emotion data). When comparing how well each measurement type, discrete or dimensional, did in explaining the underlying 27-dimensional emotion space, the team found that discrete emotion data accounts for 78% of the explainable variance while dimensional data only accounts for 61% (Cowen & Keltner, 2017). Although this suggests that discrete emotion measurements may be more powerful, one issue with this comparison is that their result ignores the fact that there are almost double the degrees of freedom in the discrete emotion data, which explains how the measurement accounts for more variance. In short, there is a natural tradeoff between breadth and parsimony that should be considered when using different emotion measures.

In emotion research there is no “gold standard” of measurement (Mauss & Robinson, 2009). For example, researchers seeking to understand the patterns of facial expressions related to basic emotions might employ the Facial Action Coding System to classify expressions into discrete emotions (Ekman, 1992a). In contrast, researchers trying to understand the neurophysiological reactions to stimuli would naturally use measures of valence and arousal to map functional segregation underlying core affect (Anders et al., 2004). In short, each method reveals something unique about the emotional landscape and this becomes even clearer when examining the study of emotion and decision-making.

Emotion and decision-making

Emotion scholars have been struggling to come up with a definition of emotion that satisfices all parties for decades, and the ensuing debates have caused the field of emotion to split into the different theories described above. Thus, is not surprising that researchers in behavioral economics and decision-making have called many different mental states “emotional”, sometimes without clear connections to specific definitions or emotion models. Examples include conscious states characterized solely by subjective feelings (Lerner et al., 2004; Liu et al., 2016; Pillutla & Murnighan, 1996), to a complex combination of physiological, neural, and facial responses (Chapman et al., 2009; Osumi & Ohira, 2009; Sanfey et al., 2003; van 't Wout et al., 2006). The range of mental states called emotions in the decision-making literature is partly due to the difficulty in conceptualizing the "essence" of emotions (Zachar, 2021). Consequently, other frameworks separate from those discussed in the emotion theories section have risen to explain affective phenomena in decision-making.

The treatment of emotion in behavioral economics and judgment and decision-making (JDM) literature has been, until recently, dominated by dual systems theories of cognition (Evans, 2009). This singular focus has limited the progress towards understanding the distributed and dynamic nature of emotion during human behavior, and instead has led to an oversimplification of the impact of emotion on choice (Van Bavel et al., 2012). To illustrate, early on, the focus of decision-making was to identify how people relied on simplified cognitive strategies called heuristics (e.g., the availability and representative heuristic) to make choices (Tversky & Kahneman, 1974). During this time, affect and emotion as a motivator was largely overlooked. It was not until the next century that the role of affect became recognized in the field of judgment and decision-making, and was promptly labeled the "affect heuristic" (Slovic et al., 2007), which was the label for how affect can be used as a mental shortcut to lead us astray. These studies combining judgment and affect would lead subsequent researchers to search for a broader framework that could explain the mechanisms governing cognitive and affective mistakes.

The popularity of the dual system—one that is quick, automatic, and reflexive, and another that is slow, deliberative and thoughtful, which roughly map onto emotion and reason, respectively (Kahneman, 2011)—led many researchers to posit that the automatic system causes errors during judgment and decision-making (Haidt, 2001; Johnson & Tversky, 1983). The litany of evidence documenting these ‘error prone choices’ was re-branded ‘affective impulses’ (Figner et al., 2009), giving many the impression that emotional responses cause people to deviate from normative or ‘rational’ choices. This is neatly demonstrated in one of the most widely used economic games, the Ultimatum Game (Güth et al., 1982). In the classic version, participants are paired up and one person is made the Proposer and endowed with an amount of money (e.g., $10). The Proposer then makes a monetary split with the other participant, the Responder, who can either accept the offer as is, or reject it. Rejecting results in both players receiving nothing (essentially, a form of costly punishment). The normative framework suggests that rejections are irrational, as receiving even a small amount of money should be considered more valuable than receiving no money. However, across numerous studies that span time and cultures, people consistently reject unfair offers (Henrich et al., 2001), a result that has been chalked up to emotion (Civai et al., 2010; van Winden et al., 2008; Yamagishi et al., 2009).

The most popular theory explaining rejections in the Ultimatum Game is the “wounded pride / spite model” (Pillutla & Murnighan, 1996). Unfair treatment evokes a negative emotional reaction--i.e., anger—that causes people to reject the offer. Evidence for specific emotional experiences like anger was only indirectly assessed by independent raters coding responses to open-ended questions. In an attempt to find a more direct link between rejection and specific emotions, subsequent studies focused largely on mood inductions through emotionally evocative videos designed to elicit discrete emotions. Compared to neutral or happy videos, both sadness (Harlé & Sanfey, 2007) and disgust (Moretti & di Pellegrino, 2010) increased rejections of unfair offers. However, there are issues with conducting single experiment mood inductions that make comparing these effects difficult. It is possible, for example, that none of these specific emotions are associated with decisions to punish in the real world. Instead, any negative feeling might be sufficient and may simply drive attention towards unfair offers (Liu et al., 2016). What these studies lack is a framework that allows for generalizing beyond discrete emotion states and their connections to punishment, which could be found by examining the underlying affective processes (e.g., valence and arousal) that change as a result of the assessment of the relative value of social choices.

One example for how a specific affective process, such as physiological arousal, is implicated in decision-making comes from multi-attribute decision tasks. In these tasks, the decision-making process can be thought of in the following components: 1) evaluate safe and risky options, 2) estimate the subjective value of each option, and 3) select one of the options. Some options are simple to evaluate, such as selecting between a sure $5 or a sure $6 which only vary on a single dimension, while other options might have multiple attributes, such as choosing to go to a cheap restaurant far away or an expensive, close restaurant. Affective processes can be involved in all stages of the decision-making process. Participants experiencing arousal re-examine information less and ignore product-describing attributes, suggesting changes in option attribute evaluation (Lewinsohn & Mano, 1993). Studies linking affective responses to risky gambles, finds that individuals who are more sensitive to losses relative to gains (i.e., more loss averse) experience higher physiological arousal on loss trials compared to gains (Sokol-Hessner et al., 2009). Using cognitive regulation strategies, participants were able to regulate their arousal to attenuate their loss aversion. Taken together, these results suggest a direct link between arousal and subjective valuation of losses, which has led some researcher to conclude that affect is synonymous with value itself (Hartley & Sokol-Hessner, 2017; Rolls & Grabenhorst, 2008).

Challenges with measuring emotion during decision-making

The current approaches to studying the role of emotion during choice suffer from two major problems: 1) demand characteristics associated with discrete emotions, and 2) generalizability of results. Focusing on the role of discrete emotions is ideal when developing theories about specific emotion states, for example the role of guilt in decision-making (Chang et al., 2011), but there are limitations to the range of emotional experiences that can be reliability elicited. While mood inductions are generally successful in evoking positive or negative moods (Fernández-Aguilar et al., 2019; Zhang et al., 2014), there is significant variability in eliciting specific emotions (Engebretson et al., 1999). If researchers opt for self-report checklists over mood inductions, it leaves open the possibility that participant's responses are merely conforming to demand characteristics. For instance, interrogating how much anger a person feels in response to unfair treatment may create an expectation that person *ought* to feel anger so no other emotional response is offered (Kassam & Mendes, 2013). Simply put, asking directed probes with a limited set of options may increase false reporting of emotional states and may miss important features of the emotional experience that are not being queried (e.g., how positive they feel). There is a need for an emotion measurement which can test discrete emotion claims without falling prey to the difficulties described above.

The other major issue with current approaches is the lack of generalizability. While some research points to the relationship between specific emotions and behaviors (e.g., action tendencies), it is hard to understand what the broader relationship is between emotion and choice. Appraisal theories provide a better framework for generalizability as they take a more dimensional approach to measuring emotion along various cognitive evaluations or appraisals (e.g., goal relevancy, safety, control, etc.). However, there is still disagreement over which set of appraisals are sufficient to capture the critical features of emotion, and interjecting multiple questions (e.g., up to 14 appraisals in recent work; Cowen & Keltner, 2017) in a decision-making paradigm is cumbersome when asking participants to report short-lasting emotional experiences. Taken together, current approaches do not provide a generalizable framework for how conscious-subjective feelings influence decisions and valuation.

Current research

In this body of work, I propose a generalizable framework for dynamically measuring emotions during decision-making paradigms. Inspired by approaches taken by Russell and colleagues (Russell et al., 1989; Tseng et al., 2014), I developed a measurement of emotion where I can continuously assess affect along the dimensions of valence (pleasant-unpleasant) and arousal (low-high). This measure is ideally suited for the questions we want to ask as it is easy to nest within behavioral economic paradigms, allowing for rapid and repeated assessment of core affect. Other checklist measures cannot deal with rapid fluctuations of affect during a social interaction, and are time consuming to implement. I argue that this is one tool in particular that can 1) mitigate many concerns over measurement, 2) sidestep the debates that have plagued the field for decades, and 3) offer a precise and mathematical framework for measuring affect in conjunction with behavioral assays. In short, this measure of emotion aims to balance parsimony, speed, and generalizability.

The first half of my research evaluates this technique for measuring emotions in real-time during social decision-making tasks. The first chapter of this thesis addresses the long-standing challenge of quantifying emotions and the traditional narrow framework in which emotion is often examined. To address this gap, I combine machine learning algorithms with a measure of affect that is agnostic to any specific emotion label. The result is a probabilistic map of emotion that is used to classify the specific emotions experienced by participants in a variety of social interactions (using the Ultimatum Game, Prisoner's Dilemma, and Public Goods Game). Results reveal that punitive and uncooperative choices are linked to a diverse array of negative, neutrally arousing emotions, such as sadness and disappointment, while only weakly linked to anger. These findings stand in contrast to the commonly held assumption that anger drives decisions to punish, defect, and freeride—thus, offering new insight into the role of emotion in motiving social choice. The second chapter extends this measure of emotion to experiences that unfold in the real-world, where the consequences of emotional responses and behaviors have tangible consequences. I tested how evoking negative and positive emotional responses through public health messages influences people's willingness to self-isolate during the COVID-19 pandemic. While results show that both types of appeals increased willingness to self-isolate when compared to a threat message, the efficacy of the prosocial message was more dependent on the magnitude of the evoked emotional response along both arousal and valence dimensions. These results imply that prosocial appeals have the potential to be associated with greater compliance if they evoke highly positive emotional responses.

In the second half of my thesis, I use a combination of behavioral, computational, and electrophysiological methods to examine how my measurement of emotion is used to assess violations of emotion expectations. I label these violations emotion prediction errors, and probe whether these emotion prediction errors influence the way people make social decisions in learning contexts. Specifically, in Chapter 3, I show across four studies that emotion and reward prediction errors have distinguishable contributions to choice, such that emotion prediction errors exert the strongest impact during decision-making. I additionally find that a choice to punish or forgive can be decoded in less than a second from an evolving emotional response, suggesting that emotions swiftly influence choice. Finally, individuals reporting significant levels of depression exhibit selective impairments in using emotion—but not reward—prediction errors. In the fourth and final chapter, I bring these findings forward to test whether there is separability at the neurobiological level. In this additional data set, I replicate the finding that emotion and reward prediction errors have independent contributions to social choice, and that these prediction errors play a critical role in trial-by-trial learning. Participants rely more on violations of emotion expectations when uncertainty about social partners is high, and being to progressively rely more on reward prediction errors as more information is gathered about how their partners behave. Different electrophysiological signatures were found to index reward, arousal, and valence PEs, respectively. Taken together, these results suggest that emotion prediction errors are a key mechanism in shaping how we make decisions and learn about social others. This generalizable framework for emotion is meant to bridge emotion theory and the decision-making field to broaden our understanding of the role of emotion during social choice.

References

Anders, S., Lotze, M., Erb, M., Grodd, W., & Birbaumer, N. (2004). Brain activity underlying emotional valence and arousal: a response-related fMRI study. *Hum Brain Mapp*, *23*(4), 200-209. https://doi.org/10.1002/hbm.20048

Arnold, M. B. (1960). *Emotion and personality*. Columbia University Press.

Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (2016). *Handbook of Emotions, Fourth Edition*. Guilford Publications. https://books.google.com/books?id=bDBqDAAAQBAJ

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560-572. https://doi.org/10.1016/j.neuron.2011.02.056

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In bad taste: evidence for the oral origins of moral disgust. *Science*, *323*(5918), 1222-1226. https://doi.org/10.1126/science.1165565

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition*, *114*(1), 89-95. https://doi.org/https://doi.org/10.1016/j.cognition.2009.09.001

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75.

Cowen, A., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900-E7909. https://doi.org/10.1073/pnas.1702247114

Cowen, A. S., & Keltner, D. (2021). Semantic Space Theory: A Computational Approach to Emotion. *Trends Cogn Sci*, *25*(2), 124-136. https://doi.org/10.1016/j.tics.2020.11.004

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227-268. http://www.jstor.org/stable/1449618

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Barrett, L. F., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., Deldin, P. J., Gotlib, I. H., & Jonides, J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychol Sci*, *23*(11), 1410-1416. https://doi.org/10.1177/0956797612444903

Du, S., Tao, Y., & Martinez Aleix, M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454-E1462. https://doi.org/10.1073/pnas.1322355111

Chapter 2

A probabilistic map of emotional experiences during competitive social interactions

Joseph Heffner & Oriel FeldmanHall
Published in Nature Communications (2022)

Introduction

Although early models of decision-making largely ignored the influence of emotion (Becker, 1976), the last few decades have detailed the central role emotions play in guiding choices (Frijda, 1986; Lerner et al., 2015; Loewenstein & Lerner, 2003; Phelps et al., 2014). Emotions are particularly important for social cognition (Keltner & Lerner, 2010; Parkinson, 1996; Tiedens & Leach, 2004), where they are known to potently shape the decisions made during interactions with other people (Rilling & Sanfey, 2010; Van Kleef et al., 2010; Zajonc, 1998). The coupling between emotion and social choice is perhaps most well-documented by research examining behaviors in competitive contexts, such as defecting during cooperative games and punishing norm transgressors. There is evidence that negatively arousing emotions—such as anger—act as the proximate mechanism motivating punishment (Andrade & Ariely, 2009; Bosman & van Winden, 2002; Carlsmith et al., 2008; Frijda et al., 1989; Moretti & di Pellegrino, 2010; Pillutla & Murnighan, 1996). For example, when reading about a moral violation (Goldberg et al., 1999; Lerner et al., 1998; Quigley & Tedeschi, 1996) or receiving an unfair offer in an economic exchange (Nelissen & Zeelenberg, 2009), anger seems to scale with willingness to punish the transgressor. Physiological and neural data corroborate this account. Enhanced physiological arousal (i.e., skin conductance and pupil dilation) is observed when punishing those who behave unfairly and defecting after experiencing unreciprocated cooperation (Ben-Shakhar et al., 2007; Decety et al., 2004; Murphy et al., 2020). This heightened negative arousal is often taken as a proxy for anger (Moretto et al., 2010; Van't Wout et al., 2006), a narrative which has also been applied to the coupling between increased neural activity and decisions to punish and defect (de Quervain et al., 2004; Rilling et al., 2004; Sanfey et al., 2003). Although less prevalent, there is some evidence that other emotions such as sadness (Harlé et al., 2012; Harlé & Sanfey, 2007), disgust (Moretti & di Pellegrino, 2010), and disappointment (Martinez & Zeelenberg, 2015; Van Doorn et al., 2012; Van Kleef & Van Lange, 2008; Wubben et al., 2009, 2011) also contribute to non-cooperative behaviors. Taken together, the prevailing account is that strong, negatively arousing emotions play a critical role in discouraging prosocial decisions to reciprocate, cooperate, or help others (Harlé et al., 2012; Sanfey, 2007).

However, the long-standing challenge of precisely quantifying nebulous emotional experiences (Dalgleish, 2004; Scherer, 2005) and the nature in which emotion is traditionally probed, leaves open the possibility that other emotions which do not neatly fit into this emotional taxonomy, might also shape social decision-making. For instance, interrogating how much anger a person feels in response to unfair treatment may artificially impose an expectation that the person ought to feel anger (Kassam & Mendes, 2013). These types of directed probes also constrain the emotional experiences to a limited set chosen by the experimenter, often without allowing participants the option to report other emotional experiences. Moreover, attributing specific emotional states to increased physiological and neural activity may falsely lead to mis-identifying the emotion actually experienced (Poldrack, 2006). Therefore, although emotion is often invoked as the lynchpin of motivated social decision-making (FeldmanHall & Chang, 2018; Lerner et al., 2015), it remains unclear which emotions drive competitive decisions to punish and defect, and which emotions govern cooperative decisions to collaborate and reciprocate.

To circumvent these issues and precisely characterize the relationship between emotion and social choice, we developed a technique for understanding emotional experiences that combines participants' affective ratings of specific emotions with the affective experiences generated during social interactions. By leveraging both supervised and unsupervised learning algorithms trained to classify emotions, we let our data-driven framework reveal the nature of this relationship, effectively reverse engineering the emotions experienced during social interactions. This agnostic, unbiased approach identifies which specific emotional experiences drive competitive and cooperative choices, including punishment of a norm transgressor, defection after experiencing unreciprocated cooperation, and free-riding at the expense of the common good.

In three experiments (N = 1,491), we embed a mathematically tractable measurement of emotion into a series of economic games: The Ultimatum Game (Experiment 1), the Prisoner's Dilemma (Experiment 2), and a Public Goods Game (Experiment 3). This emotion measure partitions emotional experiences into a broad two-dimensional affect space of valence (pleasurableness, x-axis) and arousal (activation/intensity, y-axis; Russell & Barrett, 1999), parameterized by a 500x500 pixel grid (Fig 1A). In our modified games, participants play as the second mover, reporting on this affect grid how they feel about the choices of the other player(s) (Fig 1B). For example, if their partner proposes a highly unfair offer and the participant feels extremely negatively aroused, they could move their cursor to the upper left corner of the grid, demarcating their affective experience at a specific [x, y] location (participants were never asked to self-report a discrete emotion label during any of the economic games). Participants then respond by deciding to punish, defect, or free ride—depending on the game. Critically, before playing the economic game, all participants completed an emotion classification task in which they used their memory and prior knowledge (Feldman, 1995) to place 20 labeled emotion terms (angry, surprised, happy, etc.) within the affect grid, such that each emotion is associated with a specific [x, y] coordinate (Fig. 1A). These [x, y] ratings of the labeled emotion terms are then used to train machine learning models to generate a representation of the group's 'emotion map' within the affect grid space, which enable the model to predict which labeled emotion terms are most likely to be associated with a participant's unlabeled, affective experiences reported during the economic games.

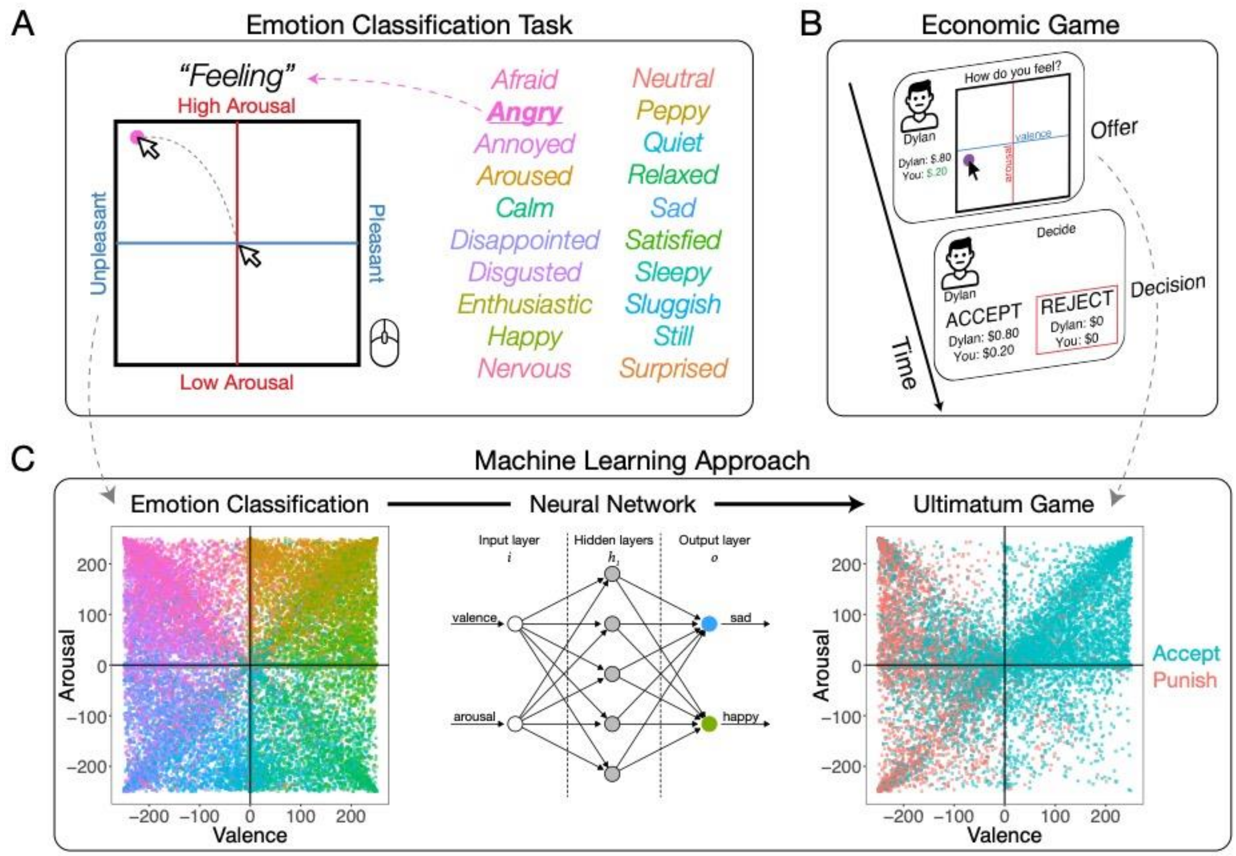


Figure 1. ***Methods A) Emotion classification task schematic.*** *Participants rated a variety of feeling labels using the arousal (vertical) and valence (horizontal) axes on a 500x500 pixel grid. The feeling labels for the emotion classification task are listed.* ***B) Economic game schematic.*** *In a second task, participants played as the second mover in a dyad or group setting, deciding whether to punish a norm transgressor (UG) or defect in a cooperation game with one (PD) or three (PGG) partners. Here we provide a schematic of Experiment 1: The Ultimatum Game. Participants are paired with a partner, receive an offer, rate how they felt about that offer, and choose to accept or reject the offer.* ***C) Machine learning approach.*** *Emotion classification*

ratings (colors correspond to emotion classification task words) were used to train a feedforward neural network, shown in a simplified schematic predicting two emotion classes (see Methods). After cross-validation, the trained neutral network was applied to the unlabeled emotion experiences in the Ultimatum Game. Ultimatum Game data is color coded by the choice to accept (blue) or reject (red).

To identify which specific emotions drive choices to punish, defect, and freeride, we trained three supervised machine learning classification algorithms (a neural network, k-nearest neighbors algorithm, and a support vector machine; Fig. 2), and one unsupervised algorithm (k-means clustering) on the valence and arousal ratings (i.e., [x, y] coordinates) provided by participants during the emotion classification task. Supervised models are evaluated on their cross-validation accuracy (see Methods), and once trained, the output of these models is the probability associated with an emotion's location in the affect grid at the population level (Fig. 2). Both the neural network (NN) and k-nearest neighbors (kNN) achieved high overall testing accuracy (NN: 35.80% and kNN: 35.97%, compared to null accuracy of 5%), while the SVM was a poor classifier (19.90%). While we chose the neural network model as our final model, results are comparable across NN and kNN. We applied the trained NN model to the unlabeled affective experiences reported during the economic games to infer an unbiased, participant-driven estimate of what emotion the person was likely feeling during the social interaction, without constraining participants' experiences (Fig. 1C).

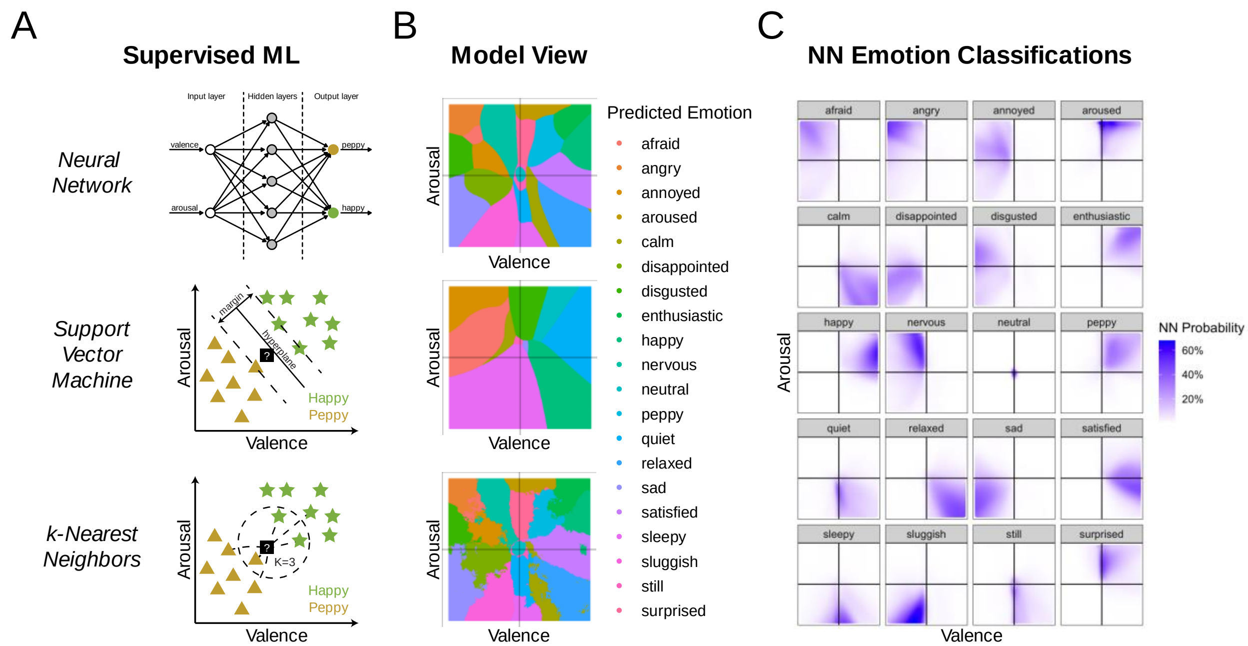


*Figure 2. **Machine learning approaches**. **A) Supervised machine learning.** Three supervised machine learning algorithms were used to train a classifier on all 20 emotions from the emotion classification data. Simplified schematics for each approach are shown demonstrating how the model classifies toy data as either "happy" or "peppy". Model-specific parameters were optimized using 10-fold cross-validation on the training dataset. **B) Model view.** Graphs show most likely emotion class for each valence-arousal location which visualizes one model's view of the emotion space. **C) Neural network emotion classifications.** Visualizes the probability of each emotion label classification for the emotion space, with darker colors indicating higher probability.*

Results

The emotions associated with decisions to punish

Participants (N = 715) played a modified Ultimatum Game (UG), where the Proposer splits a sum of money with the participant, who can then decide to accept the offer, in which case the money is split as proposed, or reject the offer, in which case neither player receives any money (a classic form of costly punishment). Offers less than 20% of the total pie are typically rejected about half the time, which punishes the transgressor for behaving unfairly (Fehr & Krajbich, 2014). In our version, participants played as the Responder (or a third party, see Methods), and reported how they felt about the Proposer's offer on the affect grid before deciding to accept or reject. Before playing in the UG, participants rated 20 feeling terms in the emotion classification task. To gain insight into the emotions experienced during the UG, we combined the emotion classification data with the affect ratings made in all three experiments using machine learning algorithms. To accomplish this, we split the emotion classification data into a training (70%) and testing set (30%), and trained a feed-forward neural network (NN) on the training data. We used cross-validation to optimize the size of a single hidden layer to prevent overfitting (see Methods). The testing set was used to validate the accuracy of the NN model by comparing the NN predicted emotion classifications, to the real emotion labels in the test data (NN achieved 35.80% accuracy, compared to a null accuracy of 5%). The resulting cross-validated NN model was applied to the unlabeled affective experiences reported in response to the Proposer's offer. Each affect rating was given a probability (or likelihood) of being classified as one of the 20 discrete labeled emotion terms (where all probabilities add up to 1). Because each affect rating was associated with a choice to accept or reject, results were averaged within participants and then across choices to gain insight into which emotions are linked to decisions to punish.

The 2D structure of emotion

Results from the emotion classification task reveal at the population level how emotions are organized in the two-dimensional valence-arousal space. For example, the contour plots of the 2D density distributions for the emotion anger reveals a tight, densely centered set of responses that fall in the high arousal, unpleasant grid space, an experience that might be described as 'rage' (Figure 3A). There is also, however, a small dense cluster of responses that are neutrally arousing and highly unpleasant, an experience probably more akin to ‘quiet anger’. These two distinct clusters of responses suggest that at the population level, there is a qualitative difference in the emotional experience labeled anger. Other emotions, such as disappointment, appear to have greater heterogeneity, reflected by responses spanning a greater swath of the affect grid, especially along the arousal dimension. This suggests that there is no single common emotional experience that neatly describes the emotion called disappointment, and even basic emotions such as anger appear to have affective variability (Figure 3B). The heterogeneity in the structure of emotions can be tested by quantifying the distribution (e.g., standard deviation or interquartile range; Barrett et al., 2001; Gruber et al., 2013) and comparing the variance of each emotion (Snedecor & Cochran, 1983). For example, many negative emotions (i.e., afraid, angry, annoyed, disgust, disappointment, and sadness) have comparable variances along the valence dimension, but most emotions’ variances differ along the arousal dimension (see Supplementary Figure 3). The distributions themselves can also be formally compared by determining whether a specific distribution is unimodal (e.g., Hartigan's dip statistic; Freeman & Dale, 2013; Hartigan & Hartigan, 1985)—which, depending on the peaked-ness of the distribution, might suggest a common emotional experience at the population level, or multimodal—which would suggest greater heterogeneity at the population level (see Supplementary Tables 1 and 2). Together, these techniques help to reveal the structure of emotion and offer insight into the heterogeneity of the human emotional experience.

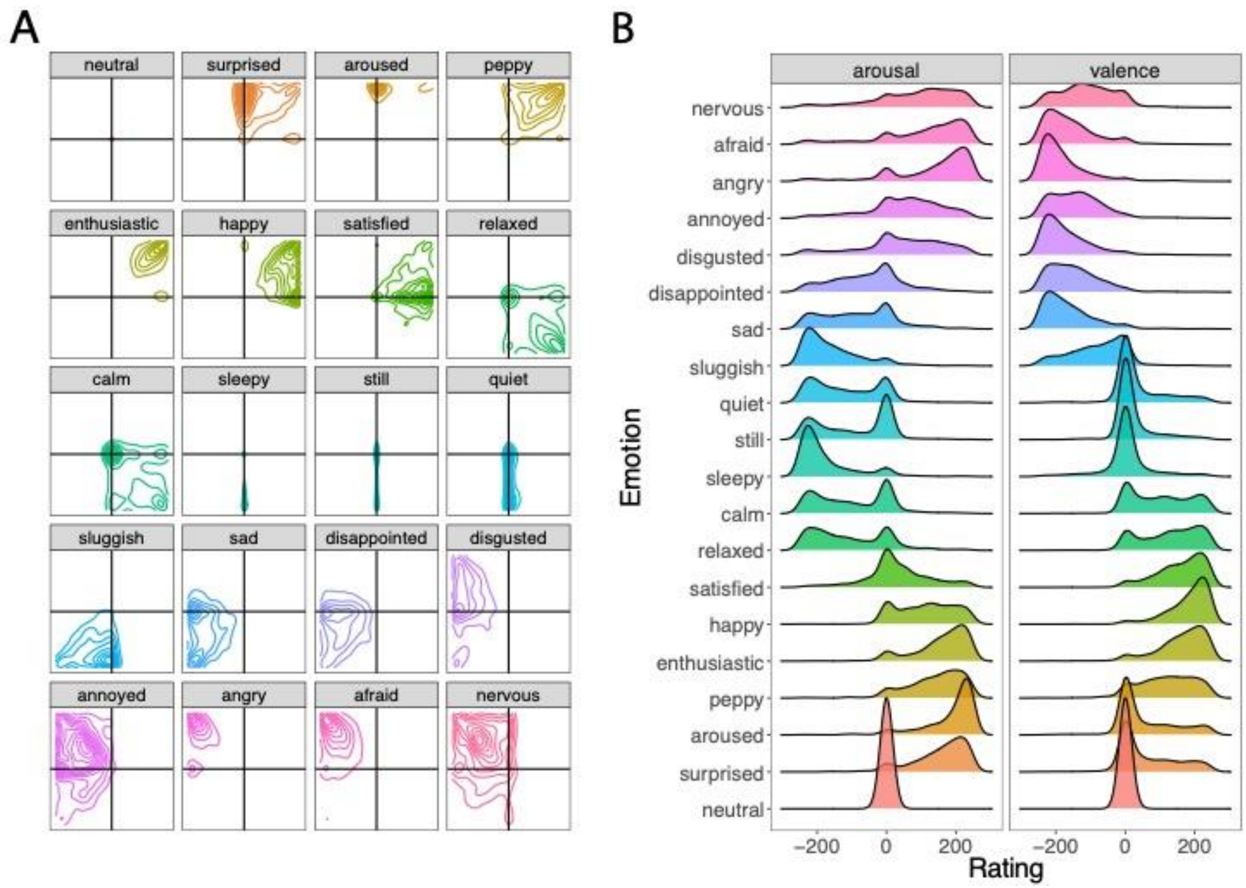


Figure 3. Emotion classification plots. A) *Two-dimensional density plots of where participants (N = 1,491) placed each emotion term in the emotion classification task. The x-axis represents valence and the y-axis represents arousal. Contour lines illustrate different levels of density at the population level, see Supplement for more detailed visualizations.* ***B) One-dimensional density plots illustrating*** *the ratings of emotion terms in the emotion classification task, plotted separately for valence and arousal.*

Sadness and disappointment are associated with punishment

The neural network trained on the emotion classification data predicts the likelihood a given emotion is experienced after receiving an offer in the UG. Contrary to popular emotion-punishment theories, our results reveal that the top three emotions associated with decisions to punish are sadness (13.47%), disappointment (12.82%), and disgust (12.54%), with anger (10.28%) identified as the 5th most likely emotion to be experienced after receiving an offer (Fig. 4A). Paired t-tests reveal, sadness ($t(558) = 4.65$, $p < .001$, $d = 0.20$), disappointment $t(558) = 4.06$, $p < .001$, $d = 0.17$, and disgust ($t(558) = 5.52$, $p < .001$, $d = 0.23$) are all significantly more likely to be associated with punishing compared to anger. When accepting the offer, the top three emotions most likely to be experienced are satisfied, happy, and peppy (Fig. 4A). To assess the degree to which a given emotion is specifically associated with a decision to punish compared to accept, we can contrast the model likelihoods for each choice (e.g., a delta measure computing the likelihood of feeling anger when punishing – likelihood of feeling anger when accepting). Higher values indicate that an emotion has a stronger association with punishment, and lower values indicate that an emotion has a stronger association with acceptance (Figure 4B). Emotions such as sadness, disgust, anger and disappointment were all more intimately linked to decisions to punish, compared to when deciding to accept.

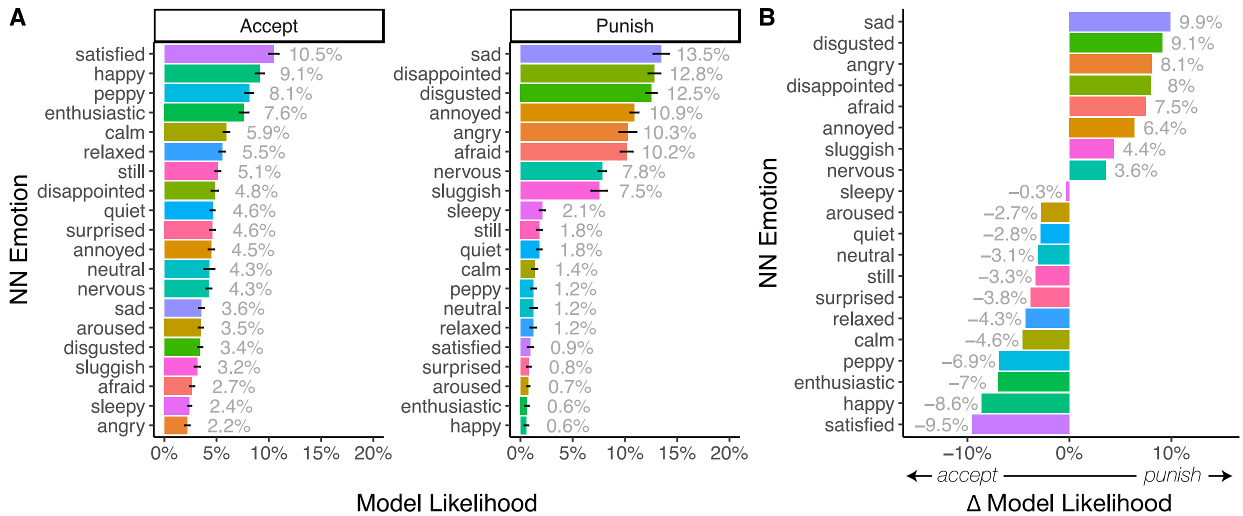


Figure 4. Neural network emotion classifications*. *A) Classifications by decisions to accept and punish. *The neural network model trained on the emotion classification data is applied to the unlabeled emotion ratings from the Ultimatum Game. Each data point is assigned a probability of each emotion class and model likelihoods are averaged within participant (N = 715) and then across choice. Bars represent mean values while error bars reflect 95% CIs.* ***B) Difference in probability of choice given an emotion.*** *Model likelihoods are averaged within participants, then across choices to reveal which emotions are more likely to be associated with decisions to accept and punish. For example, the emotion sad is more specific to decisions to punish than accept, whereas the emotion sleepy is equally associated with both choices.*

Negatively valenced, high arousal emotions less likely to predict punishment

To test the more general notion that emotions which are negatively valenced and highly arousing are commonly linked to punishment (Pillutla & Murnighan, 1996; Sanfey et al., 2003), we used an unsupervised machine learning approach called k-means clustering. The k-means procedure ignores the emotion labels from the emotion classification task and partitions the emotion space into nine equally sized clusters (we chose k = 9 because it forms a three-by-three checkered emotion space; Fig. 5A). These nine clusters are roughly equivalent to delineating all possible combinations of low, medium, and high levels of arousal and valence. The objective of the k-means algorithm is simple: While ignoring actual emotion labels, take the population-level responses from the emotion classification task to discover any underlying patterns for how the group represents each emotion (see Methods). This approach allows anger-punishment theories a fighting chance by interrogating—in an even more unbiased approach—the hypothesis that negatively valenced, highly arousing emotions (such as, anger) predict decisions to punish.

Results show that cluster 3, which corresponds to high negative valence and neutral arousal, was the most frequent affective experience associated with decisions to punish—almost twice as frequent $(\chi^2(1) = 249.68, p < .001)$ as the next predictive cluster (cluster 1: high negative valence and high arousal; Fig. 5B: note that the proportion of punish and accept choices which naturally fall into these clusters are not uniform and as such, the base rates of punishment in each cluster differ, e.g., out of all choices to punish, 11.5% fall within Cluster 3 while only 3.44% fall within Cluster 1; see Supplementary Figure 7). Examining which emotions fall into cluster 3 reveals disappointment as the most representative emotion term, whereas anger is most representative of cluster 1 (Fig. 5A). When deciding to accept the offer, cluster 5 (neutral valence and arousal; represented by the emotion term neutral) was almost twice as likely $(\chi^2(1) = 520.22, p < .001)$ as cluster 2 (neutral arousal and positive valence; represented by the emotion term satisfied). Given that we still find that the affect cluster associated with anger (cluster 1) is not representative of decisions to punish (even when accounting for base rates) suggests that our results cannot simply be explained by how people interpret specific emotion terms. Instead, these results provide converging evidence that negative valence and high arousal affective states are not the predominate motivator of punishment.

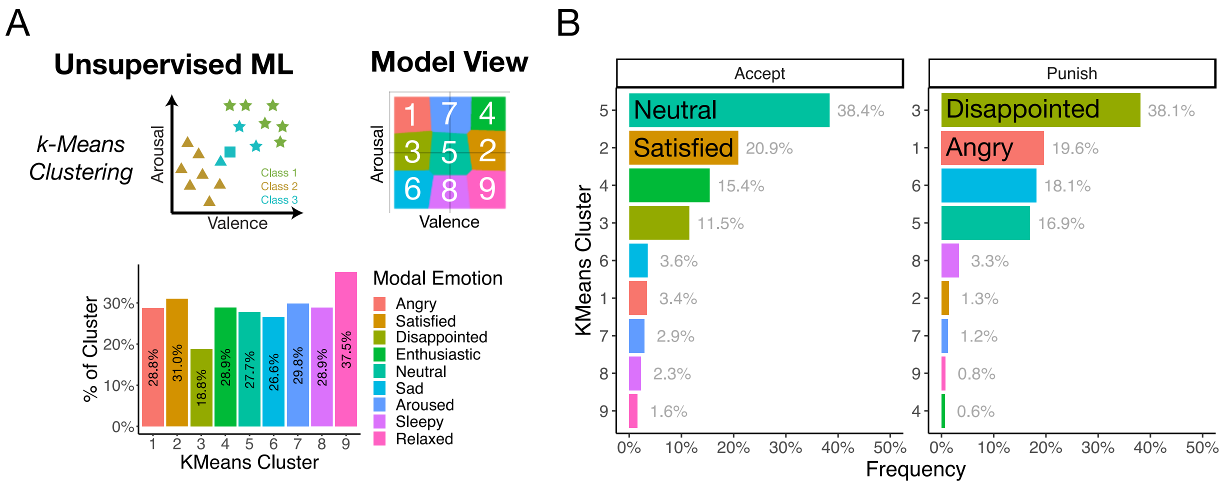


Figure 5. k-means clustering approach. A) k-means cluster schematic and interpretation. *We specified the k-means algorithm to identify nine clusters which roughly forms a three-by-three grid varying in low, medium, and high valence and arousal. Each cluster is numbered and the frequency of the modal emotion for each cluster is shown to illustrate each space of the emotion grid.* ***B) Classifications by decisions to accept and punish.*** *The k-means cluster model was applied to the unlabeled emotion ratings from the Ultimatum Game. Each data point is categorized by a single cluster and the frequency of each cluster is shown for accept and punish decisions.*

Highly unfair offers increase coupling between negatively valenced, neutrally arousing emotions and punishment

Being treated unfairly is a strong and consistent predictor of decisions to punish, especially in the UG (Camerer, 2003). The association between unfairness and punishment provides an additional, strong test of which specific emotional experiences evoked by being treated unfairly leads to punishing. Using the emotion classification likelihoods derived from the neural network, we calculated the average probability of each emotion class at every level of unfairness, given a decision to punish or accept the offer. As offers become more unfair, people experience more negative emotions including sadness, disgust, disappointment, and anger (Fig. 6). The model identified disappointment as the most likely emotion for most offer types. Only after experiencing the most unfair offer (5% of the total pie), does anger become the top emotion linked to punishment, followed by sadness and disgust (paired sample t-test comparing anger with sadness $t(537) = 0.03$, $p = 0.98$, $d = .001$) and disgust ($t(537) = 0.59$, $p = 0.55$, $d = .03$).

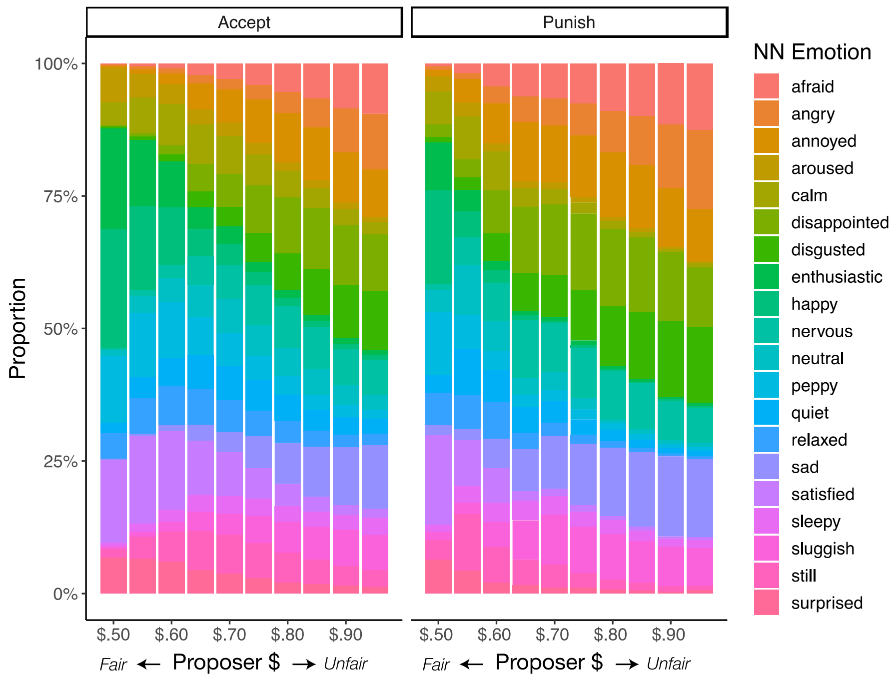


Figure 6. Neural network classifications across unfairness levels. *Unfairness indexes the amount of money kept by the Proposer (out of $1) and ranges from fair ($.50-$.50) to highly unfair offers ($.95-$.05). The proportion refers to the model likelihood of each emotion class within participants (N = 715), and then across choice (accept or punish) and each unfairness level.*

Individual-specific emotion representations

While the machine learning approach combines the data across all participants to classify affective experiences as a population-level estimate, we can also tailor our analysis to each individual's unique affective experience. To do this, we directly tied a participant's specific emotion classification responses to their reported affective experiences in the UG, which creates an individual-level estimate. For example, if a particular participant rates anger as being a negative, low arousal state (i.e., a 'quiet' anger), then this analysis would be sensitive to that participant's idiosyncratic representation of anger, which might deviate from the group's representation. To create an individual-level estimate for the emotions felt during decisions to punish, we computed the inverse Euclidean distance between a participant's affect ratings made during the UG and the coordinates reported for each emotion term in the emotion classification task (see Methods for how we converted distance to probabilities). Essentially, this allows us to capture the probability that a person's unlabeled affect ratings made during a social interaction are similar (or not) to their unique experience of a given emotion term (e.g., anger). Results reveal that at the individual level, disgust (9.41%), disappointment (8.15%), and anger (7.97%) were the most frequently experienced emotions (disgust was significantly more likely to be experienced than anger ($t(558) = 3.50$, $p < .001$, $d = 0.15$), while disappointment was not ($t(558) = 0.42$, $p = .68$, $d = 0.02$); see Supplementary Figure 9). When deciding to accept, the three emotions most likely to be experienced were neutral, happy, and satisfied. These individual-level estimates, which account for the presence of idiosyncratic emotion representations, largely align with the population level results.

The emotions associated with decisions to defect

By using machine learning algorithms to infer the emotions experienced during a social exchange, Experiment 1 found that disappointment is the most representative emotion of decisions to punish after being treated unfairly, whereas anger seems to play a much smaller role than originally believed. In Experiment 2 we wanted to further test the link between anger and social decisions in a different competitive context. Thus, we leveraged a similar experimental framework, but this time examined choices to cooperate or defect. In Experiment 2 (N = 306), participants played a modified sequential Prisoners' Dilemma (Clark & Sefton, 2001), where players decide whether to cooperate or defect with one another. In our version, both players are given \$1 and asked how much they want to contribute to a joint outcome. Any amount contributed is multiplied by 1.5 and redistributed evenly amongst both players, creating a continuous version of cooperate (contribute \$1) and defect (contribute \$0). Participants report how they feel about their partner's contribution (which ranged from \$0 - \$1, see Methods) on the affect grid before deciding how much they themselves want to contribute.

Sadness and disappointment are associated with defection

To make the analysis analogous to Experiment 1, we binned continuous contributions into decisions to defect (\$0 - \$0.49) and cooperate (\$0.50 - \$1). Results reveal that disappointment (8.83%), sadness (8.72%), and disgust (7.38%; in that order) were the most likely emotions associated with defection, all of which were more likely than anger (5.11%)—which ranked as the 9th most likely emotion. Paired t-tests reveal that disappointment ($t(278) = 8.26$, $p < .001$, $d = 0.50$), sadness ($t(278) = 6.84$, $p < .001$, $d = 0.41$), and disgust ($t(278) = 8.37$, $p < .001$, $d = 0.50$), are all significantly more likely to be associated with defection in the PD compared to anger. These results hold if we analyze choice continuously (\$0 - \$1), which only further confirms that anger ranks significantly below (8th) both sadness (1st) and disappointment (2nd; see Supplementary Figures 11 and 12). In contrast, decisions to cooperate are linked to happiness (16.32%), satisfaction (14.60%), and enthusiasm (13.88%; in that order; see Supplementary Figure 10 for full ranking). Results from the unsupervised machine learning approach which makes no assumptions about specific emotion labels per se, illustrates that the main cluster associated with defection is cluster 5 (i.e., feelings of neutral valence and neutral arousal)—which is more than five times more common than cluster 1 (i.e., feelings of high negative valence and high arousal; $\chi^2(1) = 662.88, p < .001$). The individual-level analysis dovetailed with these results, demonstrating that anger ranks as the 11th most likely emotion associated with decisions to defect (see Supplementary Figure 14 for full ranking). Together, these results provide converging evidence that anger is highly unlikely to be the main emotion motivating competitive social decision-making.

The emotions associated with decisions to freeride

To further test the generalizability of these results, in Experiment 3 (N = 470), participants played a modified sequential Public Goods Game(Varian, 1994), where participants and three players collectively decide how much to contribute to a common pool. All players were given $1 and told that any amount contributed would be multiplied by two and then redistributed evenly amongst all players. Participants reported how they felt about their partners' collective contribution on the affect grid before determining their own contribution (see Methods).

Sadness and disappointment are associated with free-riding

As in Experiment 2, we binarized continuous contributions into decisions to defect ($0 - $0.49) or cooperate ($0.50 - $1). Results reveal the top three emotions associated with defection in a group context are, sadness (10.36%), disappointment (9.63%), and sluggishness (8.38%)—with anger (4.74%) ranking 8th (see Supplementary Figure 15 for full ranking). Paired t-tests revealed that sadness ($t(442) = 12.9$, $p < .001$, $d = 0.61$), disappointment ($t(442) = 14.0$, $p < .001$, $d = 0.67$), and sluggishness ($t(442) = 7.60$, $p < .001$, $d = 0.36$) were all significantly more likely to be associated with decisions to freeride than anger. As in Experiment 2, these results hold if we analyze choice continuously ($0 - $1), illustrating that both sadness (1st) and disappointment

(2^{nd}) rank significantly above anger (7^{th}; see Supplementary Figures 16 and 17). As before, decisions to cooperate were linked to feeling happy (13.47%), enthusiastic (12.95%), and satisfied (12.06%). The unsupervised machine learning model results showed the main cluster associated with defection was again cluster 5 (associated with neutral valence and arousal), which was more than six times more likely than cluster 1 (associated with high negative valence and arousal; $\chi^2(1) = 4948.80, p < .001$). Moreover, the individual-level analysis revealed anger as the 9^{th} most likely emotion when deciding to freeride (see Supplementary Figure 19).

Discussion

Which emotions best predict decisions to punish, defect, and freeride in competitive social interactions? While prior research argues that feelings of anger play a predominant role in motivating such competitive social choices(Pillutla & Murnighan, 1996), our data-driven machine learning approach finds that other emotions, primarily disappointment and sadness, are far more likely to be associated with punishing, defecting and freeriding. Despite the intuitive appeal of negative, highly arousing emotions being linked to competitive social decision-making, we did not find good evidence that this was the case—regardless of the number of people involved or the unique social tensions of the interaction. When we account for individual variability in the representation of emotion, we find that disappointment is still more frequently experienced than anger. Even when ignoring emotion labels entirely and focusing exclusively on unlabeled affect responses, we find little evidence that highly arousing and negatively valenced feelings, an affective experience traditionally associated with the emotion anger, shape these decisions. Instead, negatively valenced and neutrally arousing emotions—experiences akin to sadness or disappointment—are the most common affective experiences associated with punishing, defecting, and freeriding.

By not forcing people to report emotions assumed to be involved with certain social choices, we observed a diverse array of emotional experiences driving competitive social decision-making. Although a forced-choice design can at times be a useful tool, it necessitates that participants' responses align with the expectations of the experimenters (Russell, 1994), and it may inflate the importance of certain emotions according to social norm theory (e.g., 'I ought to feel angry after being treated unfairly'; Russell, 2003; Russell, 2017). By using unlabeled affect measurements to probabilistically classify which emotional experiences relate to discrete social choices, our framework sidesteps these issues. Moreover, this unbiased, data driven approach reveals how unlabeled affective responses at the individual level scaffold the structure of a given emotion at the population level. For example, some people might be motivated by 'quiet anger' when punishing, while others are more motivated by an 'intense rage'. Since no singular approach can accurately describe the subjective emotional experience of all individuals (Barrett, 1998), future research should consider this type of emotion measurement in conjunction with other techniques, such as self-report or physiological measures.

Detailing how emotion interacts with decision-making is foundational for understanding the mechanisms guiding social cognition. As one prominent example, it has long been theorized—and demonstrated—that specific emotions motivate certain actions, such as a fear response leading to a tendency to flee (Damasio, 1994; Frijda, 1986; Lang & Bradley, 2010). The framework employed by traditional affect theories and emotion measurements suggests a one-to-one mapping between certain emotions and choice. Yet, leveraging a broader data-driven approach that relies less on the discretization of emotion and which makes no assumptions on the likelihood of candidate emotions experienced, enabled us to discover the wide heterogeneity of the feelings evoked during a specific set of social interactions. That is, the emotions driving punishment, defection, and freeriding are far more diverse than previously thought.

Given that we found highly negatively valenced emotions to be more closely linked with decisions to punish, defect, and freeride, our results highlight the essential role of valence in predicting social decisions. In contrast, high arousal was not as readily linked with these choices. This diverges from prior work implicating autonomic nervous system arousal as a key component motivating decisions to punish (Sanfey et al., 2003; Van't Wout et al., 2006), which suggests there might be a difference in people's awareness of their emotional processes and how the body physiologically indexes that process (Dunn et al., 2010). It may be the case that people are less able to appraise their body's physiological arousal states, which would indicate that self-reported arousal does not bear a one-to-one mapping with physiological arousal. Instead, there may be different affective roles between the conscious level associated with reporting arousal states, and the more implicit measurements garnered at the physiological level. Indeed, the relationship between physiology and subjective evaluation of emotions is not straightforward (Kreibig, 2010), as physiological arousal can be interpreted in different ways depending both on the context (Schachter & Singer, 1962), and an individual's level of introspection (Price & Hooven, 2018). Further research in this area, particularly efforts to characterize the relationship between physiological and reported arousal (LeDoux & Brown, 2017), can help better characterize this interactive, dynamic process, and how it might bias choice.

Although the two-dimensional affect structure of valence and arousal is central to theories of emotion (Barrett et al., 2006; Lewis & Haviland-Jones, 2000; Russell, 1980; Russell, 2003) and grounded in the body's neurobiological system (Posner et al., 2005), some aspects of emotional experience may not be properly captured by a two dimensional valence-arousal model. For example, anger and disgust are both associated with unpleasant, high arousal affect, but are conceptually linked to different behavioral responses: Anger is typically associated with approach-motivated responses and disgust with avoidance-motivated responses (Carver, 2006; Carver & Harmon-Jones, 2009; Curtis et al., 2011; Oaten et al., 2009). While the two dimensional valence-arousal space offers a quick measurement of affect and is thus useful for fast, repeated decision-making paradigms, adding in more dimensions (e.g., approach/avoid, anticipated effort, control; Cowen & Keltner, 2017; Skerry & Saxe, 2015; Smith & Ellsworth, 1985), could be helpful in better characterizing the structure of the emotional space, and should improve classification accuracy by separating similar emotions along other dimensions.

Although the general link between emotion and choice is established in the field of human social cognition, less is understood about the relationship between the specific emotions guiding social choices. Our results suggest that framing anger as a motivating force may mischaracterize the nature of the relationship. While we observed large variability in the emotions driving competitive social choices, emotions characterized by high, negatively valenced, but muted arousal appeared to play the most central role in driving decisions to punish, defect, and freeride. By combining the continuous, dimensional structure of affect with discrete emotion states, our novel analytic approach allowed us to identify the likelihood that a particular emotion drives a social choice, while also revealing the heterogeneity of emotions experienced during these interactions. As the disciplines of emotion and behavioral economics advance, we can build on this progress to further characterize how emotions guide decisions in a variety of contexts.

Methods

Participants

Across three experiments, participants (N = 1,820) completed an emotion classification task followed by one of three economic games: An Ultimatum Game (UG; Experiment 1 N = 906), a Prisoners' Dilemma (PD; Experiment 2 N = 395); or a four-person Public Goods Game (PGG; Experiment 3 N = 519). Using preregistered criteria based on existing work (Heffner et al., 2021), we excluded participants (N=329 total; UG exclusion = 191; PD exclusion = 89; PGG exclusion = 49) who rated neutral outside of a 100x100-pixel square in the center (i.e., participants were instructed to rate "neutral" in the center of the dARM). The final sample for the UG was N = 715 (320 Females, mean age = 34.4 +/- 10.1), the final sample for the PD was N = 306 (131 Females, mean age = 35.5 +/- 11.2), and the final sample for the PGG was N = 470 (238 Females, mean age = 33.0 +/- 10.5). Thus, the final sample across all three experiments was N = 1,491 (689 Females, mean age = 34.2 +/- 10.5). Participants were recruited from Amazon Mechanical Turk and received monetary compensation and provided informed consent in a manner approved by Brown University's Institutional Review Board under protocol 1607001555.

Procedure

In the Emotion Classification Task, participants rated a series of feeling words on an affect grid varying on valence and arousal (Fig. 1), modified from the traditional affect grid (Russell et al., 1989; Russell & Gobet, 2012; Tseng et al., 2014). A person who is rating anger might, for example, place their cursor at the top left corner of the grid, indicating high arousal and negative valence (participants are free to report any subjective interpretation of anger, such as, a quiet anger). The 20 feeling terms (neutral, surprised, aroused, peppy, enthusiastic, happy, satisfied, relaxed, calm, sleepy, still, quiet, sluggish, sad, disappointed, disgusted, annoyed, angry, afraid, nervous) were selected from past research to represent the octants of the emotion circumplex space evenly(Feldman, 1995) or because they have been implicated in decisions to punish and defect during social interactions (Liu et al., 2016). Participants then completed one of three economic games which were structured in a similar way. Finally, participants completed a series of individual difference questionnaires that were not analyzed in this research.

Ultimatum game

Participants completed 20 one-shot Ultimatum Games as either the Responder (N = 543) or a third-party (N = 172). Since Responders and third-parties' affective responses were not significantly different, we collapsed across role (see Supplement). After receiving the offer, participants reported how they felt on the affect grid before deciding to accept or reject. Participants received an even distribution of offers ranging from fair ($.50, $.50) to highly unfair ($.95, $.05). Unfairness was operationalized numerically according to the amount of money kept by the partner. Participants engaged with new partners on each round. The data from the UG was part of a dataset collected for a previously published study (Heffner et al., 2021).

Prisoners' dilemma

Participants were paired with a new partner (denoted by a face and name) on each round. Both players were given $1, which they could use to contribute to a collective pot. Any amount contributed was multiplied by 1.5 and redistributed evenly between the pair. The tension lies between making more money by defecting at the expense of the other player's monetary gain. Our PD version was structured in a sequential manner, so that emotion could be measured during the game, and to compare it to the Ultimatum Game data. Participants decided how much they wanted to contribute in $.10 increments. Participants completed 22 one-shot Prisoners' Dilemma rounds, where partners contributions ranged from defection ($0) to full ($1) cooperation.

Public goods game

The Public Goods Game was similar to Experiment 2 except participants were paired with three other players (players were denoted by an anonymous Amazon Mechanical Turk IDs). The players were each given $1, which they could use to contribute to the common pot. Once contributed, the money was doubled and redistributed evenly between all four players. A player could make more money at the expense of the outcomes of others by freeriding while the others pay into the common pot. After participants were informed of the collective amount contributed by the other players, they reported how they felt on the affect grid about the collective contributions. Then they decided how much to contribute themselves (up to $1, in $.10 increments). Participants completed 62 sequential one-shot Public Goods Games with an even distribution of partner contributions ranging from none ($0) to full ($3 total contribution).

Machine learning models

In the Emotion Classification Task participants rated a variety of emotion words on the 500x500 affect grid. Participants rated 20 emotions, not all of which were related to punishing/accepting an unfair offer or freeriding amongst a group (e.g., "peppy"). This data was used to develop the supervised machine learning classifiers. We trained three supervised machine learning classification algorithms (Fig. 2) using 10-fold cross-validation on a 70-30 train-test split of the emotion classification task, and one unsupervised machine learning algorithm. Within the training data, we use 10-fold cross-validation which randomly splits the training data into 10-subsets. One subset is reversed for validation while the model is trained on the other subsets according to the model's specific algorithm. The model is tested on the reserved subset and an accuracy score is recorded. In this classification problem, accuracy is the proportion of correct classifications over the total classifications. A null accuracy would be 5% because this represents a model which simply suggests at each classification. Because the emotion classification dataset is balanced, meaning each of the emotion classes is equally represented, accuracy is a good metric for evaluating the model. The process is repeated until each of the 10 subsets have served as the validation set and the cross-validation accuracy is the average of these recorded accuracies.

Neural network

Neural networks are a nonlinear statistical model and a popular choice for complex classification problems. We used a feed-forward single-hidden-layer neural network (Venables & Ripley, 2002), which had two input nodes for valence and arousal, a single hidden layer with a variable number of hidden nodes, and 20 output nodes representing the 20 emotions from the classification task. The number of nodes in the hidden layer was determined by cross-validation and the final model contained 27 nodes in the hidden layer, with a decay of 0.035, which were fully connected to the input and output nodes.

Support vector machine

Although support vector machines are typically used for binary classification, they can be a useful tool for multi-class classifications. The support vector classifier constructs a linear boundary in a large, transformed version of the feature space (in this case, valence and arousal ratings) by defining an optimal hyperplane that separates two classes. For our multi-class problem, we used a “one-against-one” approach or pairwise classification method which constructs $\binom{k}{2}$ classifiers where each one is trained on data from two classes(Karatzoglou et al., 2004). Because our classes are not perfectly separable (i.e., they overlap and misclassifications exist within a margin of the hyperplane), we use cross-validation to determine the correct “Cost” (C) parameter, which defines the weight of how much of the data inside the margin of error contributes to the overall error. The final SVM model used a cost parameter of C = 0.01.

k-Nearest neighbors

Another popular method for classification of data with a low number of dimensions is the K-nearest neighbors (KNN) classifier. The KNN classifier is a non-parametric approach that computes the conditional probability of the data class for any data point. This is done by comparing a data point to the class of data points in close proximity. The classifier chooses a neighborhood size, represented by the parameter K, and estimates the conditional probability for emotion cluster j as the proportion of data in the neighborhood set, whose class is also j. The K

parameter and neighborhood size was determined through cross-validation, and the final KNN model used a neighborhood size of k = 175.

Model selection

Both the NN and KNN reached similar levels of overall testing accuracy (NN: 35.80%, KNN: 35.97%) and kappa (NN: 32.42, KNN: 32.60), while the SVM fit relatively poorly (SVM accuracy: 19.90%, SVM kappa: 15.68). While the KNN had marginally more accurate classifications, it also had undesirable properties, such as sporadic emotion islands (e.g., a tiny, isolated pocket of an emotion surrounded by a larger emotion cluster). The NN model achieved similar accuracy and had smooth emotion boundaries between classifications. For these reasons, we selected NN as the final model, but note that both models had comparable results. After training the NN model and validating its accuracy through cross-validation, we applied the trained neural network to the unlabeled emotion data generated from the affect grid in the Ultimatum Game, Prisoners' Dilemma, and Public Goods Game. We generated a likelihood of this emotion data being classified in each of the 20 emotion categories. That is, each emotion rating across all trials of the behavioral economic games is given a probability for each of the 20 possible emotion classes, which together add up to 1. This approach allows us to identify which emotion states are likely to drive punishment, without potentially biasing the result in favor of any specific emotion (e.g., specifically asking "how angry do you feel?").

K-Means clustering

K-means clustering is an unsupervised machine learning method for finding clusters in a set of unlabeled data. Although the emotion classification data was labeled, we can strip each emotion from its label for the purposes of finding optimal emotion clusters which vary on valence and arousal intensity. Given a desired number of cluster centers, the k-means procedure randomly defines the initial center of the cluster and then iteratively moves the center of that cluster to minimize the total within-cluster variance(Hastie et al., 2001). We chose 9 clusters because it formed a rough three-by-three checkered emotion space representing high, medium, and low valence and arousal combinations.

Euclidean distance measurements

Although the Euclidean distance measurement used in the individual difference analyses are not formal machine learning models, they represent another way to classify emotions that are unique to each participant. We calculate the Euclidean distance between the 20 emotion terms in the emotion classification task and the unlabeled affective reports during each trial of the economic games. We convert Euclidean distance into a numeric probability using inverse distance weighting, where the probability is the inverse of the Euclidean distance for a given emotion over the sum of inverse Euclidean distances of all emotions. Smaller Euclidean distances indicate a higher probability that the unlabeled affective experience is similar to the valence and arousal rating of the emotion term (e.g., anger, disgust, etc.) for each participant separately.

References

Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. *109*(1), 1-8. https://doi.org/10.1016/j.obhdp.2009.02.003

Barrett, L. F. (1998). Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition and Emotion*, *12*(4), 579-599. https://doi.org/10.1080/026999398379574

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, *15*(6), 713-724. https://doi.org/10.1080/02699930143000239

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2006). The Experience of Emotion. *Annual Review of Psychology*, *58*(1), 373-403. https://doi.org/10.1146/annurev.psych.58.110405.085709

Becker, G. (1976). *The economic approach to human behavior*. University of Chicago Press.

Ben-Shakhar, G., Bornstein, G., Hopfensitz, A., & van Widen, F. (2007). Reciprocity and emotions in bargaining using physiological and self-report measures. *Journal of Economic Psychology*, *28*(3), 314-323. https://doi.org/10.1016/j.joep.2007.02.005

Bosman, R., & van Winden, F. (2002). Emotional Hazard in a Power-to-Take Experiment. *The Economic Journal*, *112*(476), 147-169. http://www.jstor.org/stable/798435

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316-1324. https://doi.org/10.1037/a0012165

Carver, C. S. (2006). Approach, Avoidance, and the Self-Regulation of Affect and Action. *Motivation and Emotion*, *30*(2), 105-110. https://doi.org/10.1007/s11031-006-9044-7

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological bulletin*, *135*(2), 183-204. https://doi.org/10.1037/a0013965

Clark, K., & Sefton, M. (2001). The Sequential Prisoner's Dilemma: Evidence on Reciprocation. *The Economic Journal*, *111*(468), 51-68. https://doi.org/doi:10.1111/1468-0297.00588

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900. https://doi.org/10.1073/pnas.1702247114

Curtis, V., de Barra, M., & Aunger, R. (2011). Disgust as an adaptive system for disease avoidance behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1563), 389-401. https://doi.org/10.1098/rstb.2010.0117

Dalgleish, T. (2004). The emotional brain. *Nature Reviews Neuroscience*, *5*(7), 583-589. https://doi.org/10.1038/nrn1432

Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain*. G.P. Putnam.

de Quervain, D. J. F., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science*, *305*(5688), 1254.

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage*, *23*(2), 744-751.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., Cusack, R., Lawrence, A. D., & Dalgleish, T. (2010). Listening to Your Heart: How Interoception Shapes Emotion Experience and Intuitive Decision Making. *Psychol Sci*, *21*(12), 1835-1844. https://doi.org/10.1177/0956797610389191

Fehr, E., & Krajbich, I. (2014). Chapter 11 - Social Preferences and the Brain. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics (Second Edition)* (pp. 193-218). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-416008-8.00011-5

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*(1), 153-166. https://doi.org/10.1037/0022-3514.69.1.153

FeldmanHall, O., & Chang, L. J. (2018). Chapter 14 - Social Learning: Emotions Aid in Optimizing Goal-Directed Social Behavior. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-Directed Decision Making* (pp. 309-330). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-812098-9.00014-0

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behav Res Methods*, *45*(1), 83-97. https://doi.org/10.3758/s13428-012-0225-x

Frijda, N. H. (1986). *The emotions*. Editions de la Maison des Sciences de l'Homme.

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212-228. https://doi.org/10.1037/0022-3514.57.2.212

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: the psychology of the intuitive prosecutor [https://doi.org/10.1002/(SICI)1099-0992(199908/09)29:5/6<781::AID-EJSP960>3.0.CO;2-3]. *European Journal of Social*

Psychology, *29*(5-6), 781-795. https://doi.org/https://doi.org/10.1002/(SICI)1099-0992(199908/09)29:5/6<781::AID-EJSP960>3.0.CO;2-3

Gruber, J., Kogan, A., Quoidbach, J., & Mauss, I. B. (2013). Happiness is best kept stable: positive emotion variability is associated with poorer psychological health. *Emotion (Washington, D.C.)*, *13*(1), 1-6. https://doi.org/10.1037/a0030262

Harlé, K. M., Chang, L. J., van 't Wout, M., & Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: A mediating role of the anterior insula. *Neuroimage*, *61*(1), 32-40. https://doi.org/https://doi.org/10.1016/j.neuroimage.2012.02.027

Harlé, K. M., & Sanfey, A. G. (2007). Incidental sadness biases social economic decisions in the Ultimatum Game. *Emotion (Washington, D.C.)*, *7*(4), 876-881. https://doi.org/10.1037/1528-3542.7.4.876

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, *13*(1), 70-84. https://doi.org/10.1214/aos/1176346577

Hastie, T., Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behaviour. *Nature Human Behaviour*, *5*(10), 1391-1401. https://doi.org/10.1038/s41562-021-01213-6

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1-20. https://doi.org/http://dx.doi.org/10.18637/jss.v011.i09

Kassam, K. S., & Mendes, W. B. (2013). The effects of measuring emotion: physiological reactions to emotional situations depend on whether someone is asking. *PLoS One*, *8*(7), e64959-e64959. https://doi.org/10.1371/journal.pone.0064959

Keltner, D., & Lerner, J. S. (2010). *Emotion* [doi:10.1002/9780470561119.socpsy001009]. John Wiley & Sons, Inc.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394-421. https://doi.org/https://doi.org/10.1016/j.biopsycho.2010.03.010

Lang, P. J., & Bradley, M. M. (2010). Emotion and the motivational brain. *Biol Psychol*, *84*(3), 437-450. https://doi.org/10.1016/j.biopsycho.2009.10.007

LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences*, *114*(10), E2016. https://doi.org/10.1073/pnas.1619316114

Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin*, *24*(6), 563-574. https://doi.org/10.1177/0146167298246001

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799-823. https://doi.org/10.1146/annurev-psych-010213-115043

Lewis, M., & Haviland-Jones, J. M. (2000). *Handbook of emotions, 2nd ed*. Guilford Press.

Liu, C., Chai, J. W., & Yu, R. (2016). Negative incidental emotions augment fairness sensitivity [Article]. *Scientific Reports*, *6*, 24892. https://doi.org/10.1038/srep24892

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In *Handbook of affective sciences.* (pp. 619-642). Oxford University Press.

Martinez, L. F., & Zeelenberg, M. (2015). *Trust me (or not): Regret and disappointment in experimental economic games* [doi:10.1037/dec0000025]. Educational Publishing Foundation.

Moretti, L., & di Pellegrino, G. (2010). Disgust selectively modulates reciprocal fairness in economic interactions [doi:10.1037/a0017826]. *Emotion (Washington, D.C.)*, *10*(2), 169-180. https://doi.org/10.1037/a0017826

Moretto, G., Làdavas, E., Mattioli, F., & Di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of cognitive neuroscience*, *22*(8), 1888-1899.

Murphy, D. A. J., Harmer, C. J., Browning, M., & Pulcu, E. (2020). Affective biases encoded by the central arousal systems dynamically modulate inequality aversion in human interpersonal negotiations. *bioRxiv*, 826529. https://doi.org/10.1101/826529

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger and guilt and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*(7), 543-553.

Oaten, M., Stevenson, R. J., & Case, T. I. (2009). Disgust as a disease-avoidance mechanism. *Psychological bulletin*, *135*(2), 303-321. https://doi.org/10.1037/a0014823

Parkinson, B. (1996). Emotions are social [https://doi.org/10.1111/j.2044-8295.1996.tb02615.x]. *British Journal of Psychology*, *87*(4), 663-683. https://doi.org/https://doi.org/10.1111/j.2044-8295.1996.tb02615.x

Phelps, E. A., Lempert, K. M., & Sokol-Hessner, P. (2014). Emotion and Decision Making: Multiple Modulatory Neural Circuits. *Annual review of neuroscience*, *37*(1), 263-287. https://doi.org/10.1146/annurev-neuro-071013-014119

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, *68*(3), 208-224.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*, *10*(2), 59-63. https://doi.org/10.1016/j.tics.2005.12.004

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715-734. https://doi.org/10.1017/S0954579405050340

Price, C. J., & Hooven, C. (2018). Interoceptive Awareness Skills for Emotion Regulation: Theory and Approach of Mindful Awareness in Body-Oriented Therapy (MABT). *Frontiers in Psychology*, *9*, 798-798. https://doi.org/10.3389/fpsyg.2018.00798

Quigley, B. M., & Tedeschi, J. T. (1996). Mediating Effects of Blame Attributions on Feelings of Anger. *Personality and Social Psychology Bulletin*, *22*(12), 1280-1288. https://doi.org/10.1177/01461672962212008

Rilling, J. K., & Sanfey, A. G. (2010). The Neuroscience of Social Decision-Making. *Annual Review of Psychology*, *62*(1), 23-48. https://doi.org/10.1146/annurev.psych.121208.131647

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage*, *22*(4), 1694-1703. https://doi.org/https://doi.org/10.1016/j.neuroimage.2004.04.015

Russell, J. A. (1980). A circumplex model of affect [doi:10.1037/h0077714]. *39*, 1161-1178. https://doi.org/10.1037/h0077714

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol Bull*, *115*(1), 102-141. https://doi.org/10.1037/0033-2909.115.1.102

Russell, J. A. (2003). Core affect and the psychological construction of emotion [doi:10.1037/0033-295X.110.1.145]. *Psychological Review*, *110*(1), 145-172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A. (2017). Mixed Emotions Viewed from the Psychological Constructionist Perspective. *Emotion Review*, *9*(2), 111-117. https://doi.org/10.1177/1754073916639658

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805-819. https://doi.org/10.1037/0022-3514.76.5.805

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493-502. https://doi.org/10.1037/0022-3514.57.3.493

Russell, Y. I., & Gobet, F. (2012). Sinuosity and the Affect Grid: A Method for Adjusting Repeated Mood Scores. *Perceptual and Motor Skills*, *114*(1), 125-136. https://doi.org/10.2466/03.28.PMS.114.1.125-136

Sanfey, A. G. (2007). Social Decision-Making: Insights from Game Theory and Neuroscience. *Science*, *318*(5850), 598. https://doi.org/10.1126/science.1142996

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*(5626), 1755-1758. https://doi.org/10.1126/science.1082976

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379-399. https://doi.org/10.1037/h0046234

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695-729. https://doi.org/10.1177/0539018405058216

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Curr Biol*, *25*(15), 1945-1954. https://doi.org/10.1016/j.cub.2015.06.009

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *J Pers Soc Psychol*, *48*(4), 813-838.

Snedecor, G. W., & Cochran, W. G. (1983). *Statistical Methods* (6th ed.). Oxford and IBH.

Tiedens, L. Z., & Leach, C. W. (2004). *The Social Life of Emotions*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511819568

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., Colibazzi, T., Algermissen, M., Chiang, I. C., Russell, J. A., & Peterson, B. S. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *J Autism Dev Disord*, *44*(6), 1332-1346. https://doi.org/10.1007/s10803-013-1993-6

Van Doorn, E. A., Heerdink, M. W., & Van Kleef, G. A. (2012). Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment. *Cognition and Emotion*, *26*(3), 442-461. https://doi.org/10.1080/02699931.2011.648174

Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). Chapter 2 - An Interpersonal Approach to Emotion in Social Decision Making: The Emotions as Social Information Model. In *Advances in experimental social psychology* (Vol. 42, pp. 45-96). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2601(10)42002-X

Van Kleef, G. A., & Van Lange, P. A. M. (2008). What Other's Disappointment May Do to Selfish People: Emotion and Social Value Orientation in a Negotiation Context. *Personality and Social Psychology Bulletin*, *34*(8), 1084-1095. https://doi.org/10.1177/0146167208318402

Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental brain research*, *169*(4), 564-568.

Varian, H. R. (1994). Sequential contributions to public goods. *Journal of Public Economics*, *53*(2), 165-186. https://doi.org/https://doi.org/10.1016/0047-2727(94)90019-1

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). Springer.

Wubben, M. J. J., Cremer, D. D., & Dijk, E. v. (2009). How emotion communication guides reciprocity: Establishing cooperation through disappointment and anger. *Journal of Experimental Social Psychology*, *45*(4), 987-990. https://doi.org/https://doi.org/10.1016/j.jesp.2009.04.010

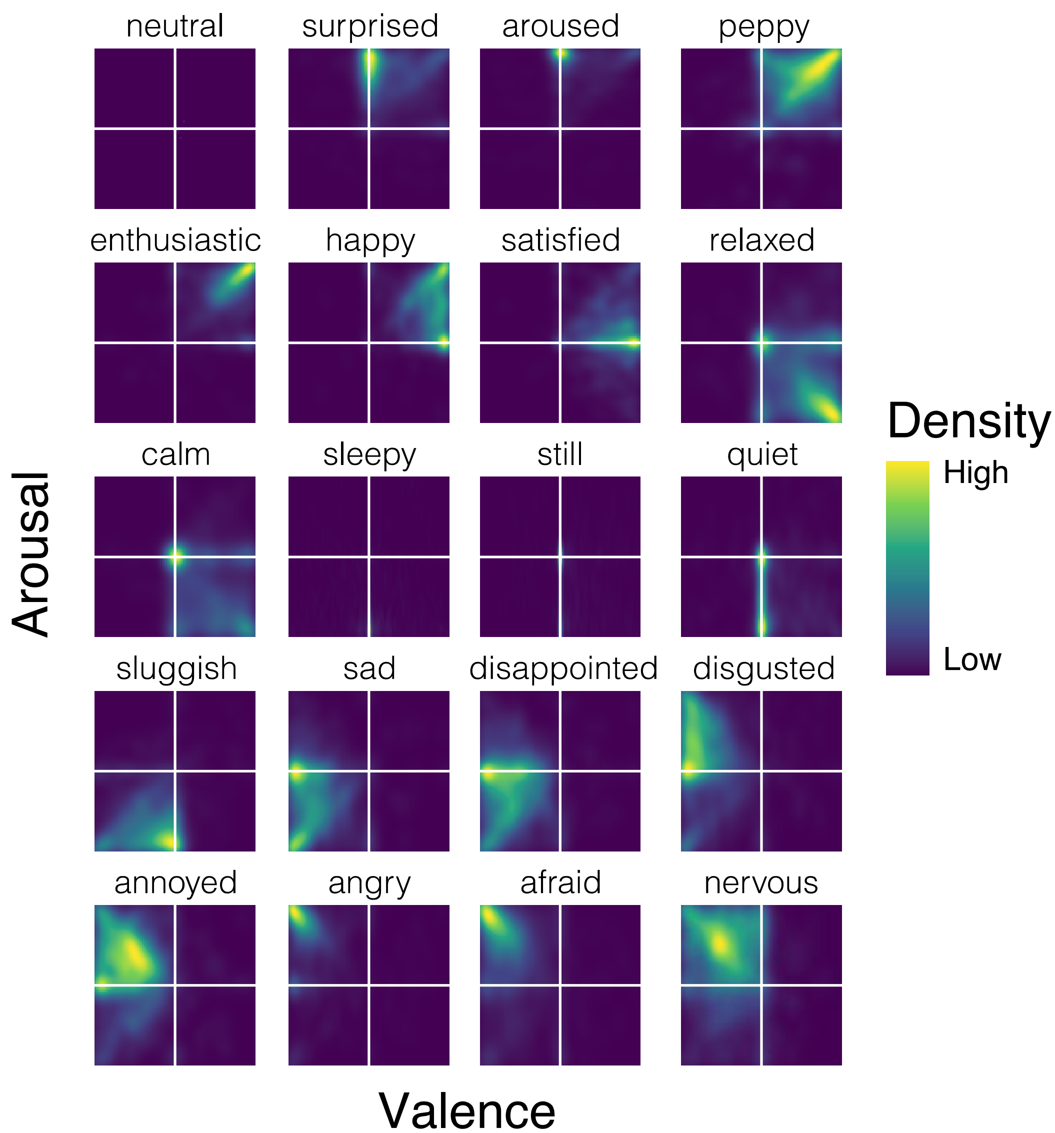
Wubben, M. J. J., Cremer, D. D., & Dijk, E. v. (2011). The communication of anger and disappointment helps to establish cooperation through indirect reciprocity. *Journal of Economic Psychology*, *32*(3), 489-501. https://doi.org/https://doi.org/10.1016/j.joep.2011.03.016

Zajonc, R. B. (1998). *Emotions*McGraw-Hill.

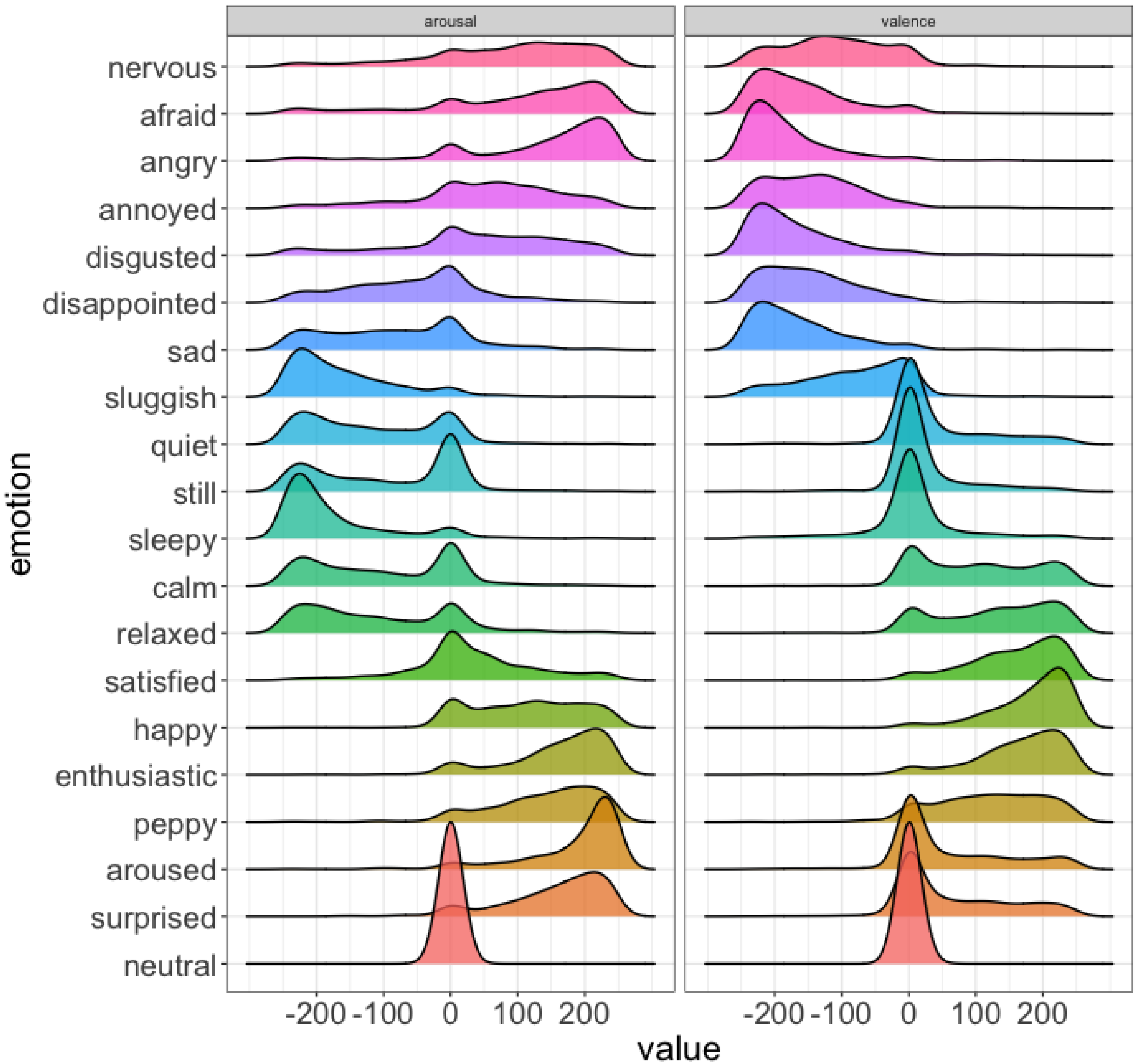
Supplementary information

1 Emotion Classification Task

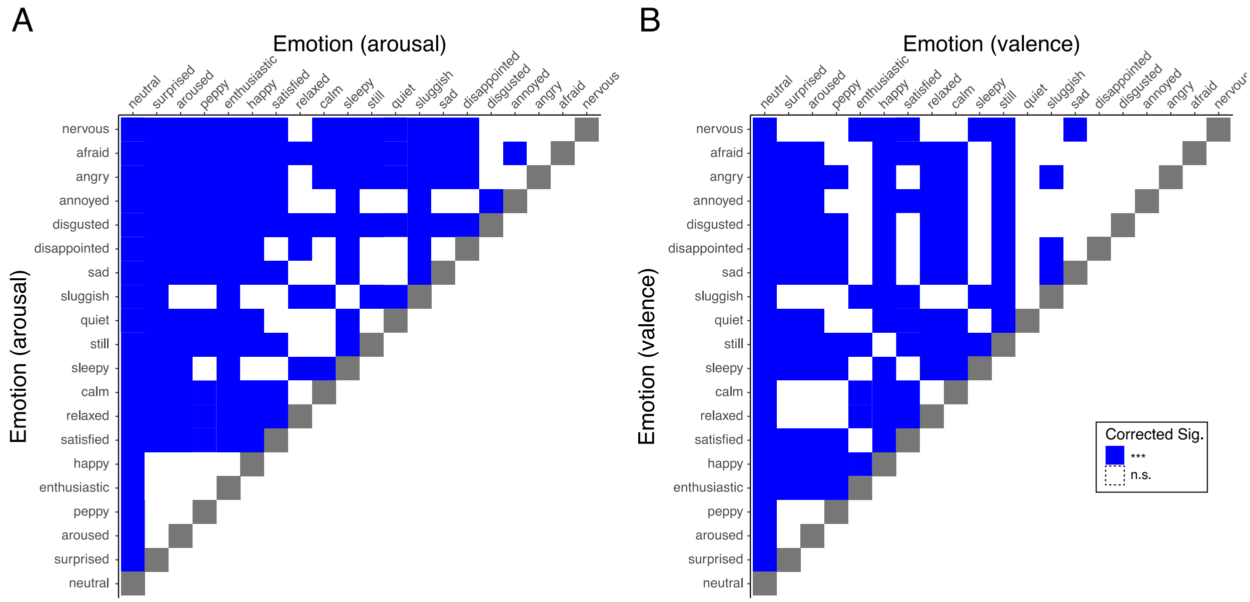
The variability associated with affective ratings from the emotion classification task can be visualized in two-dimensions or one-dimension, depending on the researcher's goal. We visualized the two-dimensional density of each emotion using contour lines in the manuscript but here visualize density using color, similar to a heatmap (Supplementary Figure 1). Because the absolute level of density is not the same across all emotions, the color mapping is the relative density within a specific emotion category. It is likely that the highest density regions reflect the canonical emotion response associated with each category, but these also reveal nuances which may be meaningful to specific emotion categories depending on theoretical points. These differences can be meaningfully quantified, as visualized in one-dimensional density plots (Supplementary Figure 2). For example, comparing variability of across affective dimensions might be important for emotions thought to be relatively homogenous. We compared the standard deviations of valence and arousal ratings between all emotions using a F-test and visualized the Bonferroni corrected significant results (Supplementary Figure 3). Another way to analyze these distributions is to see whether participants may rely on different interpretations of the emotion word (e.g., relaxed), which would result in a multimodal distribution (e.g., two modes at the qualitatively different interpretations). Researchers have used several measures to distinguish between unimodality and multimodality, but one particularly robust metric is known as the Hartigan's dip statistic (HDS; Freeman & Dale, 2013; Hartigan & Hartigan, 1985)). For example, in the case of the valence ratings for the emotion word "relaxed", results show that the distribution is significantly different from a unimodal distribution (dip statistic D = 0.0650, $p < .001$). If we assume that this distribution is bimodal, then one possible explanation for the distribution is that participants are reflecting on two distinct experiences of the emotion "relaxed" – some participants experienced this emotion as being neither pleasant nor unpleasant (neutral) while others rated this experience as highly pleasant. Supplementary Tables 1 and 2 contain uncorrected HDS tests for all emotions except neutral (because neutral is bound according to the exclusion criteria) on both the arousal and valence dimensions. These analyzes help reveal the structure of emotions on the affective dimensions and demonstrate the heterogeneity of emotional experiences.



Supplementary Figure 1. 2D Density Plots from the Emotion Classification Task. *The x-axis represents the valence dimension, the y-axis represents the arousal dimension, and color represents the relative density of valence-arousal ratings (yellow = high density, blue = low density). Density levels are relative to each emotion class.*



Supplementary Figure 2. 1D Density Plots from the Emotion Classification Task. *The x-axis represents the affect rating on either the valence dimension or the arousal dimension. The y-axis represents the density of the affect ratings, which vary for each emotion class and affect measurement. Density plots were generated with a scale of 3 and a bandwidth of 18.*



Supplementary Figure 3. Variance comparisons with F-tests. *Pairwise variance ratio tests were performed for all emotions on the arousal (A) and valence (B) dimensions. Blue tiles indicate significant variance differences after Bonferroni correction with arousal and valence comparisons.*

Emotion Label	Valence Distribution Type	Valence HDS
Afraid	Multimodal	0.015*
Angry	Unimodal	0.01
Annoyed	Multimodal	0.014*
Aroused	Multimodal	0.052***
Calm	Multimodal	0.026***
Disappointed	Unimodal	0.008
Disgusted	Unimodal	0.009
Enthusiastic	Multimodal	0.015*
Happy	Unimodal	0.012
Nervous	Multimodal	0.03***
Peppy	Multimodal	0.026***
Quiet	Multimodal	0.056***
Relaxed	Multimodal	0.055***
Sad	Unimodal	0.01
Satisfied	Unimodal	0.012
Sleepy	Multimodal	0.054***
Sluggish	Multimodal	0.02***
Still	Multimodal	0.07***
Surprised	Multimodal	0.042***

Supplementary Table 1. Multimodality in valence experiences by emotion categories. *Evidence for a multimodal distribution was determined through Hartigan's dip statistic (HDS). HDS is a null-hypothesis test, therefore a significant result indicates the rejection of a unimodal distribution. The distribution column indicates whether the valence distribution is unimodal or multimodal according to the HDS. All p-values are uncorrected for multiple comparisons.*

Emotion Label	Arousal Distribution Type	Arousal HDS
Afraid	Multimodal	0.028***
Angry	Multimodal	0.039***

Annoyed	Multimodal	0.018**
Aroused	Unimodal	0.013
Calm	Multimodal	0.061***
Disappointed	Multimodal	0.015*
Disgusted	Multimodal	0.015*
Enthusiastic	Multimodal	0.025***
Happy	Multimodal	0.036***
Nervous	Multimodal	0.024***
Neutral	Multimodal	0.098***
Peppy	Multimodal	0.018**
Quiet	Multimodal	0.066***
Relaxed	Multimodal	0.065***
Sad	Multimodal	0.016**
Satisfied	Multimodal	0.024***
Sleepy	Multimodal	0.022***
Sluggish	Multimodal	0.016**
Still	Multimodal	0.067***
Surprised	Multimodal	0.02***

Supplementary Table 2. Multimodality in arousal experiences by emotion categories. *Evidence for a multimodal distribution was determined through Hartigan's dip statistic (HDS). HDS is a null-hypothesis test, therefore a significant result indicates the rejection of a unimodal distribution. The distribution column indicates whether the arousal distribution is unimodal or multimodal according to the HDS. All p-values are uncorrected for multiple comparisons.*

2 Experiment 1: Ultimatum Game

2.1 Responders and Third Parties Respond Similarly

Participants either played the UG as the Responder or a non-vested third party. Using a logistic mixed-effects model, we examined whether affective reactions to offers depended on the participants' role. Results showed a main effect of role, such that third parties punished more than Responders, but no significant interaction between valence and role or arousal and role (Supplementary Table 3). Thus, despite the difference in their roles, all participants responded similarly on valence and arousal to unfair offers.

Responders and Third Parties Respond Similarly

	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	-3.80	0.34	-11.12	**<0.001**
Valence	-5.34	0.36	-14.80	**<0.001**
Arousal	-0.06	0.41	-0.16	0.874
Role [Self]	-1.22	0.45	-2.74	**0.006**
Valence:Arousal	-0.41	0.34	-1.23	0.220
Valence:Role [Self]	0.38	0.45	0.85	0.396
Arousal:Role [Self]	0.32	0.46	0.68	0.495
Valence:Arousal:Role [Self]	0.32	0.36	0.89	0.373
N_{sub}	364			

Supplementary Table 3. Responders and Third Parties Respond Similarly*. Valence and arousal have been scaled, but not mean-centered, as the 0 point refers to the meaningful instance when affect is neutral. Role is operationalized as self or responder (0) and third party (1). The mixed-effects model included a random intercept and random slopes for valence, arousal, and their interaction per subject and conducts two-sided t-tests for each coefficient without correction for multiple comparison.*

2.2 Affective Experiences Predict Punishment

Decisions in the Ultimatum Game (UG) may involve a policy decision (e.g., if offer below 20% of the total, reject) and here we test whether affective experiences explain decisions to punish above and beyond such a policy decision. We used mixed-effects logistic regressions to test two models of decision-making in the UG: one which models choices to punish as a function of unfairness (Supplementary Equation 1), and one which models choices to punish as a function of unfairness interacting with affective experiences (valence and arousal separately). Supplementary

Equation 1 captures the intuition that people use a policy decision (e.g., reject if below 20% of the total pie) as the slope of unfairness (β_1) will represent the strength of the policy decision while the intercept (β_0) will represent the point of indifference where participants are 50% likely to punish. (vs accept). Supplementary Equation 2 captures the intuition that emotional affective experiences improve predicting decisions to punish and potentially adjust sensitivity to unfairness.

$$punish \sim \beta_0 + \beta_1 unfairness \tag{1}$$

$$punish \sim \beta_0 + \beta_1 valence + \beta_2 arousal + \beta_3 unfairness + \beta_4 valence * unfairness + \beta_5 arousal * unfairness \tag{2}$$

We used likelihood ratio tests to compare these nested models and infer which model better explains decisions to punish. Likelihood ratio tests reveal that the model which includes affective experiences (Supplementary Equation 2) significantly improves the fit compared to the model which only included unfairness ($\chi^2(11) = 489.21,\ p < .001$). Results from Supplementary Equation 2 show significant marginal effects of unfairness and valence such that higher unfairness increases the probability of rejection while more positive valence decreases the probability (Supplementary Table 4). In short, a simpler model based purely on a decision policy of amount of money offered does a worse job of explaining behavior compared to a model that includes affective experiences.

Ultimatum Game Affect Model of Punishment

	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	-4.71	0.28	-16.61	**<0.001**
Unfairness	3.85	0.26	14.95	**<0.001**
Valence	-2.28	0.22	-10.31	**<0.001**
Arousal	-0.00	0.18	-0.02	0.982
Unfairness:Valence	-0.13	0.15	-0.91	0.363
Unfairness:Arousal	0.19	0.13	1.48	0.138
Valence:Arousal	-0.09	0.14	-0.62	0.535
Unfairness:Valence:Arousal	0.20	0.11	1.88	0.060
N_{sub}	715			

Supplementary Table 4. Ultimatum Game Affect Model of Punishment. *Valence and arousal have been scaled, but not mean-centered, as the 0 point refers to the meaningful instance when affect is neutral. Unfairness is operationalized as the amount of money kept by Player B and has been normalized (scaled and mean-centered). The mixed-effects model included a random intercept and random slopes for unfairness, valence, and arousal per subject and conducts two-sided t-tests for each coefficient without correction for multiple comparison.*

We also examined how valence and arousal combine to predict decisions to punish. Using a mixed-effect regression, we tested whether the interaction between valence and arousal predicts decisions to punish using the following fixed-effects formulation (Supplementary Equation 3):

$$punish \sim \beta_0 + \beta_1 valence + \beta_2 arousal + \beta_3 valence * arousal \qquad 3$$

Results from this mixed-effect regression show that while arousal on its own does not predict decisions to punish, there is a significant interaction between valence and arousal (Supplementary Table 5). This suggests that the joint combination of valence and arousal are critical in understanding how emotions relate to punishment. Given the variability in arousal ratings associated with decisions to punish, we also tested the possibility that arousal might predict decisions to punish in a non-linear way. We added a quadratic term for arousal to Supplementary Equation 3 and results show a significant non-linear effect of arousal, such that increases in the quadratic value of arousal strongly predicts punishment (Supplementary Table 6). Supplementary Figure 4 visualizes the main effect of arousal in this non-linear model, illustrating how the probability of punishing increases at the two extremes of the arousal scale. Furthermore, the visualized interaction between valence and arousal (Supplementary Figure 5) reveals how the relationship between valence and punishment is significantly attenuated when arousal is neutral (arousal = 0 SD) compared to when arousal is extremely low (arousal = -2 SD) or extremely high (arousal = +2 SD).

Interaction between Valence and Arousal Predicts Punishment

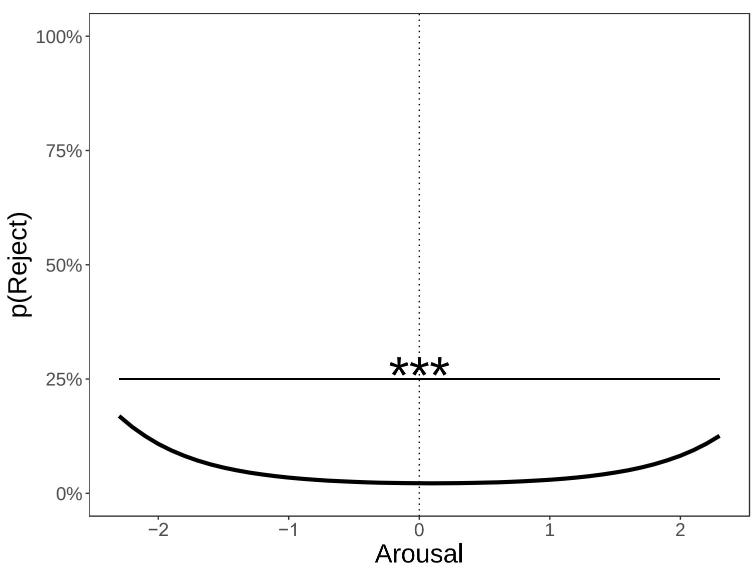
	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	-3.88	0.18	-21.64	**<0.001**
Valence	-4.51	0.18	-25.51	**<0.001**
Arousal	-0.23	0.19	-1.22	0.222
Valence:Arousal	-0.38	0.16	-2.42	**0.015**
N_{sub}	715			

Supplementary Table 5. Interaction Between Valence and Arousal Predicts Punishment. *Valence and arousal have been scaled, but not mean-centered, as the 0 point refers to the meaningful instance when affect is neutral. The mixed-effects model included a random intercept and random slopes for valence, arousal, and their interaction per subject and conducts two-sided t-tests for each coefficient without correction for multiple comparison.*

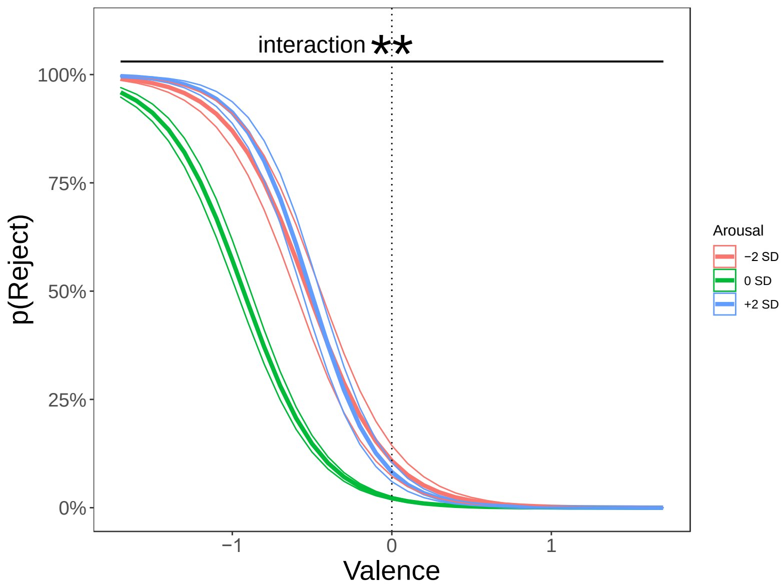
Non-linear Effects of Arousal Predicting Punishment

	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	-3.43	0.16	-20.89	**<0.001**
Valence	-4.20	0.17	-24.16	**<0.001**
Arousal(1)	7.33	11.91	0.62	0.538
Arousal(2)	66.66	10.05	6.63	**<0.001**
Valence:Arousal(1)	-24.75	9.36	-2.64	**0.008**
Valence:Arousal(2)	-12.80	8.92	-1.44	0.151
N_{sub}	715			

Supplementary Table 6. Non-linear Effects of Arousal Predicting Punishment. *Valence and arousal have been scaled, but not mean-centered, as the 0 point refers to the meaningful instance when affect is neutral. Orthogonal polynomials were used to test for non-linear effects of arousal and the degree of the polynomial is presented in paratheses (e.g., (1), (2)). The mixed-effects model included a random intercept and random slopes for valence and the orthogonal polynomial terms for arousal per subject and conducts two-sided t-tests for each coefficient without correction for multiple comparison.*



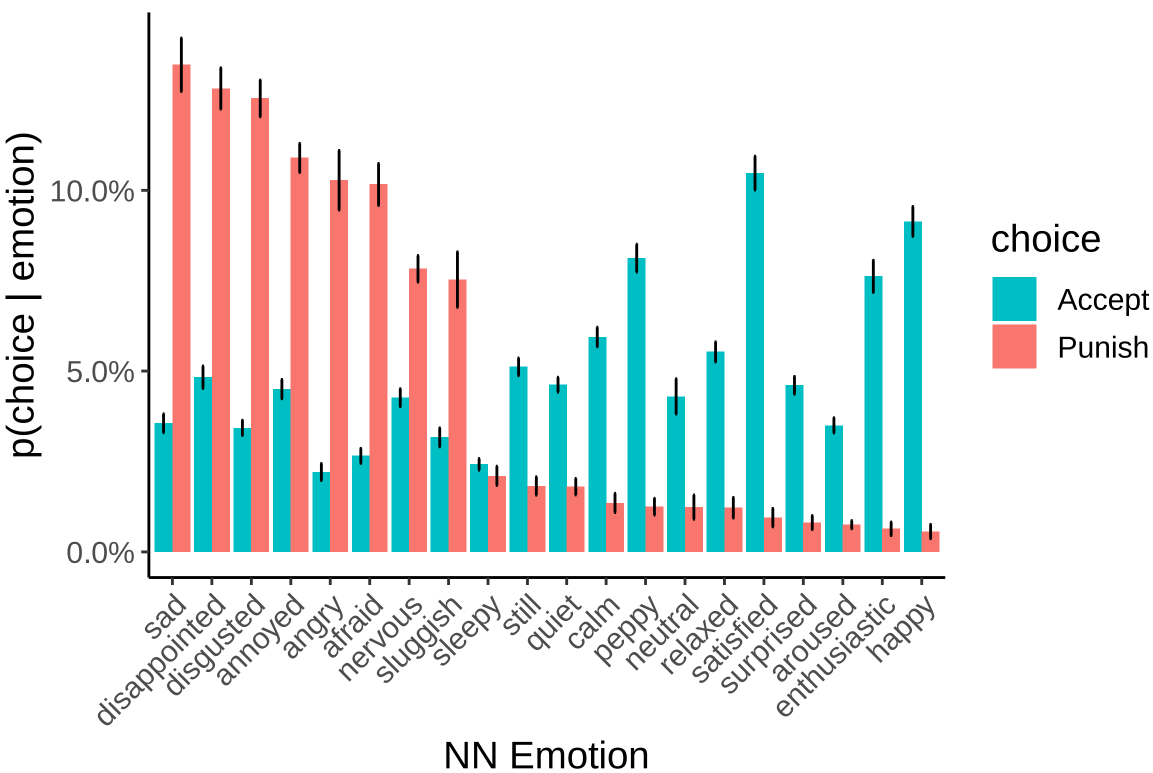
Supplementary Figure 4. Main Effect of Arousal from Supplementary Table 6. *The predicted regression line for the main effect of arousal from Table 3 is plotted for continuous values of arousal when valence equals 0. Valence and arousal were scaled, but not mean-centered, as the 0 point meaningfully refers to the case when affect is neutral. Error bars are +/- 1 standard errors of the mean.*



Supplementary Figure 5. Interaction Between Valence and Arousal from Supplementary Table 6. *Predicted regression lines from Table 3 are plotted for continuous values of valence and binned values of arousal. For visualization purposes arousal has been binned to be -2, 0, or +2 standard deviation, although the regression used the continuous value. Valence and arousal were scaled, but not mean-centered, as the 0 point meaningfully refers to the case when affect is neutral. Error bars are +/- 1 standard errors of the mean.*

2.3 Neural Network Results

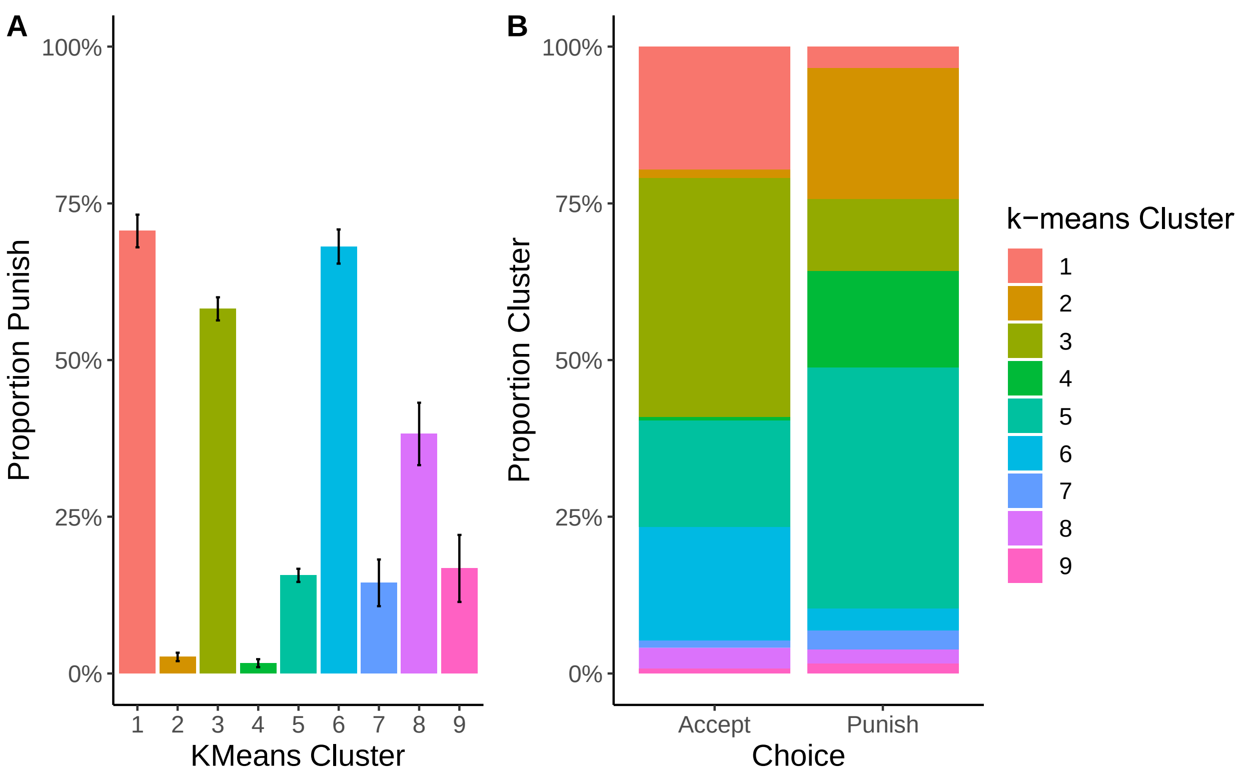
We applied the trained neural network to the affective experiences in the Ultimatum Game to generate model likelihoods for each emotion category for decisions to punish and accept. While the information from this analysis is displayed in the manuscript (Figure 4), we visualized it a different way that highlights the conditional probabilities of each choice given a specific emotion (i.e., *p(punish* | *emotion)*; Supplementary Figure 6). Using two-sample t-tests, we compared the model likelihoods associated with punishment decisions and accept decisions for each emotion (going from left to right emotions in Supplementary Figure 6). Results from this analysis show that punishment decisions are significantly more likely than accept decisions when the emotion is sadness, disappointed, disgust, annoyed, anger, afraid, nervous, or sluggish (all P's < .001). Conversely, accept decisions are significantly more likely than punish decisions when the emotion is still, quiet, calm, peppy, neutral, relaxed, satisfied, surprised, aroused, enthusiastic, or happy (all P's < .001). For the emotion sleepy, accept is significantly more likely than punish, although the effect is much weaker than the other emotions ($t(942.97) = 2.021$, $p = .04$).



Supplementary Figure 6. Conditional probabilities of choice by neural network emotion classifications. *The neural network model trained on the emotion classification data was applied to the unlabeled affect ratings from the Ultimatum Game. Each data point was assigned a probability of each emotion class and these were averaged within, then across participants (N = 715), and across choices. Emotions are organized by the conditional* *probability of choice given emotion for each emotion class. Data are presented as mean values while errors bars reflect 95% CIs.*

2.4 k-means Clustering Results

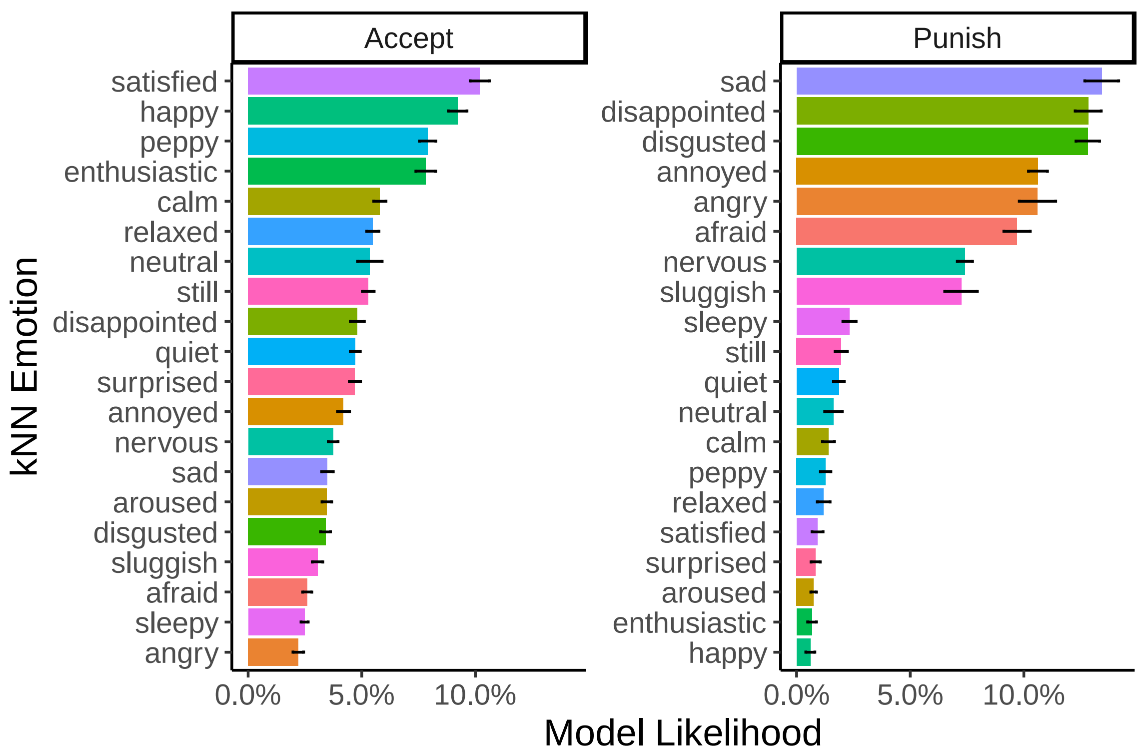
We used a k-means clustering algorithm to classify affective experiences in the Ultimatum Game into one of nine distinct clusters representing low, medium, and high valence and arousal areas. We analyzed and visualized the proportion of punish decisions within each cluster and the proportion of each decision per cluster (Supplementary Figure 7). It is important to distinguish the sensitivity of a cluster to a particular choice (Supplementary Figure 7A) from the representation of that cluster inside of all decisions to punish or accept (Supplementary Figure 7B). For example, Clusters 1 and 3 contain a high percentage of punishment decisions within each cluster respectively (Cluster 1: 70.6%, Cluster 3: 58.2%; Figure S6A). However, Cluster 1 and 3 only represent 3.4% and 11.5%, respectively, of all decisions to punish (Figure S6B). In other words, although Cluster 1 is comprised of mostly punishment decisions, the base rate of punishment decisions falling into Cluster is very low.



Supplementary Figure 7. k-means clustering proportions. A) Proportion of punish decisions by clusters. *Affective experiences in the Ultimatum Game were clustered into 9 clusters using k-means clustering. Bars represent the proportion of punishment decisions averaged within participants (N = 715) and then within each cluster. Data are* *presented as mean values while error bars reflect 95% CIs.* ***B) Proportion of clusters by choices****. Stacked bars represent the proportion of clusters which fall into accept or punish decisions.*

2.5 kNN Classification Results

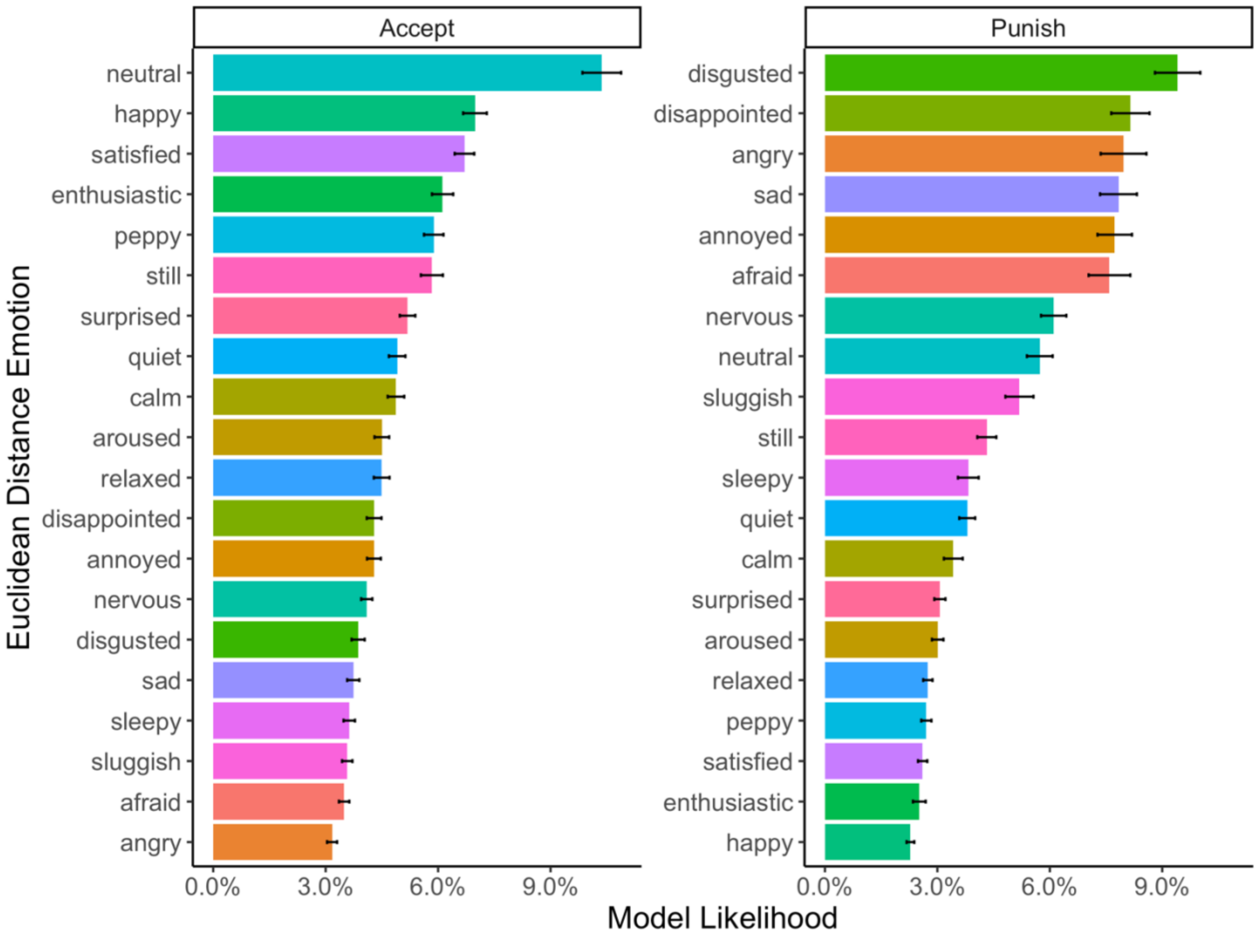
Although we used the neural network algorithm as the final classification model in the manuscript, the kNN classification algorithm accuracy was very similar to the neural network. Here, we examine the average model likelihoods associated with each emotion class and decisions to punish using the kNN algorithm. Results show that the ranking of emotions is virtually identical to the neural network results presented in the manuscript (Supplementary Figure 8).



Supplementary Figure 8. k-nearest neighbors emotion classifications by decisions to accept and punish*. The k-nearest neighbors model trained on the emotion classification data was applied to the unlabeled affect ratings from the Ultimatum Game. Each data point was assigned a probability of each emotion class and model likelihoods were averaged within participant (N = 715) and then across choice. Data are presented as mean values while error bars reflect 95% CIs.*

2.6 Euclidean Distance Analysis

We also developed an emotion classifier that is unique to each participant by calculating the Euclidean distance between all 20 emotion labels and the affective experience during each trial of the UG. We convert Euclidean distance into a probability using inverse distance weighting, where probability is the inverse of the Euclidean distance for a specific emotion over the sum of inverse Euclidean distances of all emotions. This way smaller Euclidean distances indicate a higher probability that the affective experience matched the valence and arousal rating of the emotion class for each participant separately. Results show that disgust, disappointment, and anger are the top three emotions associated with punishing (Supplementary Figure 9).

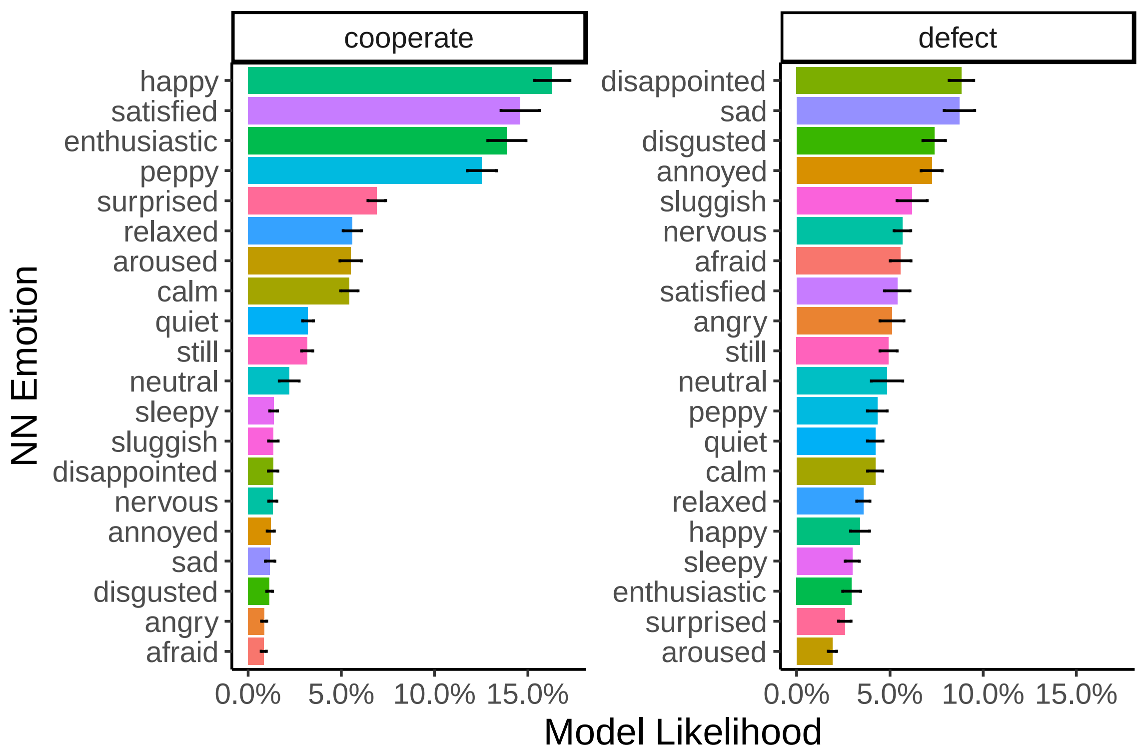


Supplementary Figure 9. Euclidean distance classifier for the UG. *Inverse distance weighting was used to calculate the probability of each emotion class for all unlabeled emotion ratings in the Ultimatum Game. This classification was done separately for each participant (N = 715), and model likelihoods were averaged within participant and then across participants. Data are presented as mean values while error bars reflect 95% CIs.*

3 Experiment 2: Prisoner's Dilemma

3.1 Discrete Contribution Analysis

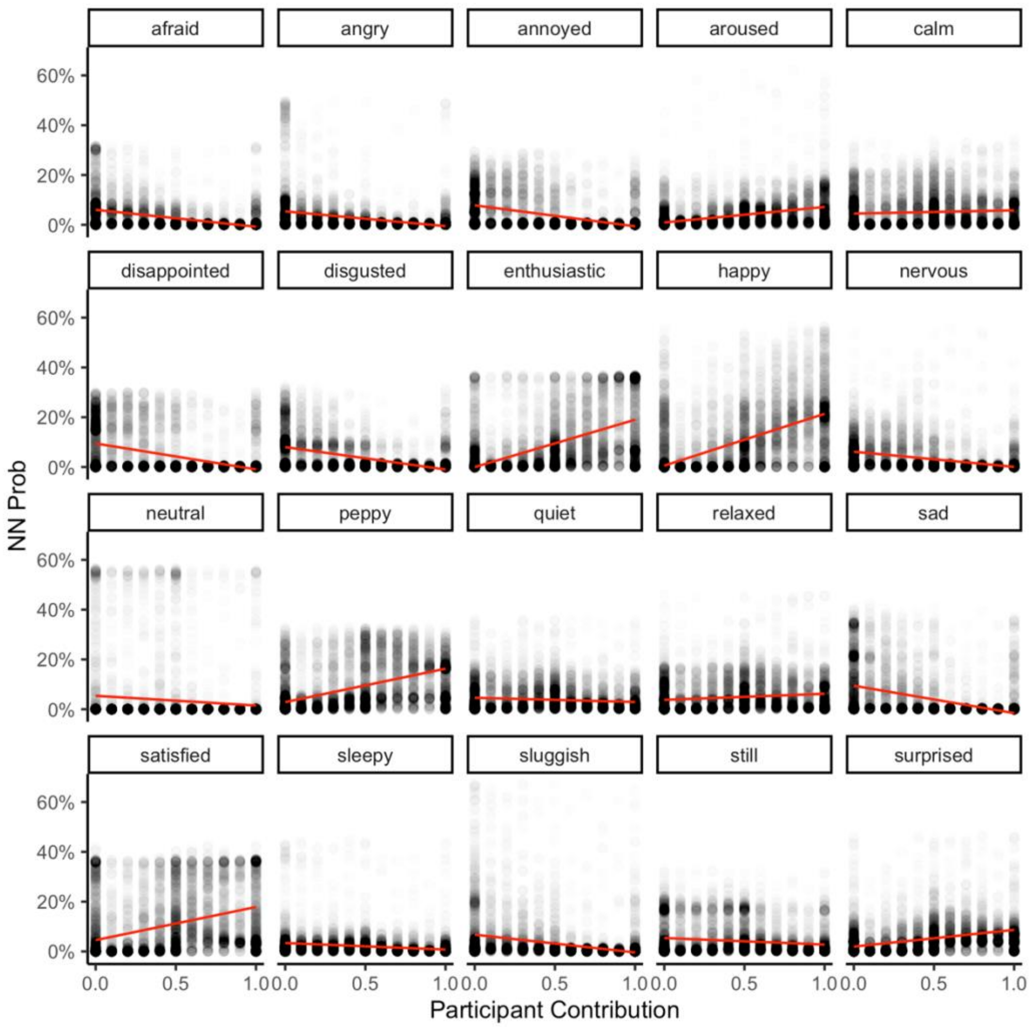
We binned continuous contributions ($0 - $1) made in the Prisoner's Dilemma (PD) into decisions to defect ($0 - $0.49) and cooperate ($0.50 - $1), and classified emotions experienced in the PD using the trained neural network from the emotion classification task. Results reveal that the top three emotions associated with decisions to defect are disappointment (8.83%), sadness (8.72%), and disgust (7.38%), with anger (5.11%) being identified as the 9th most likely emotion to be experienced in the Prisoner's Dilemma (Supplementary Figure 10). Paired t-tests revealed that disappointment ($t(278) = 8.26$, $p < .001$, $d = .50$), sadness ($t(278) = 6.84$, $p < .001$, $d = .41$), and disgust ($t(278) = 8.37$, $p < .001$, $d = .50$) were all significantly more likely to be associated with punishment than anger. When cooperating, the top three emotions most likely to be experienced were happiness, satisfaction, and enthusiasm.



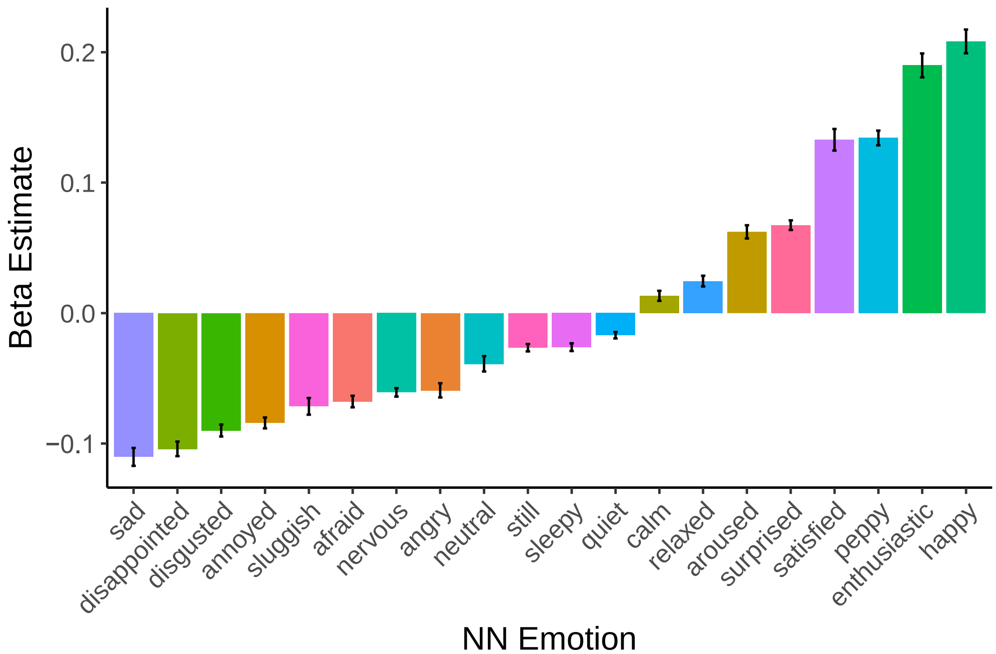
Suppelementary Figure 10. Classifications by decisions to cooperate and defect. *The neural network model trained on the emotion classification data was applied to the unlabeled emotion ratings from the Prisoner's Dilemma. Each data point was assigned a probability of each emotion class, and model likelihoods were averaged within participant (N = 306) and then across participants. Data are presented as mean values while error bars reflect 95% CIs.*

3.2 Continuous Contribution Analysis

We investigated how the probability of any specific emotion classification changed if we model continuous contributions ($0 - $1) rather than discrete contributions. We applied the trained neural network reported in the manuscript on the unlabeled emotion experiences in the PD which results in a classification probability for each of the twenty emotion terms. We ran separate mixed-effects linear regressions for each emotion term, using the participant's contribution to predict the neural network probability for that emotion term (Supplementary Figure 11). This analysis allows for more granular changes as participants increase their contributions in the PD. Examining the slopes of these regressions indexes the strength of the relationship between contributions in the PD and the probability of a specific emotion being experienced (Supplementary Figure 12). Stronger slopes indicate a stronger relationship between that emotion and contributions in the PD. For example, if monetary contributions in the PD are tightly coupled with feelings of anger, this analysis should demonstrate a strong slope between contributions and the probability of the emotion being classified as anger. Results show that compared to other emotions (i.e., sad, disappointed, and disgusted), anger was not as sensitive to changes in contribution levels. This demonstrates that even when taking the continuous nature of the contributions into account, anger is not a representative emotion associated with defection.



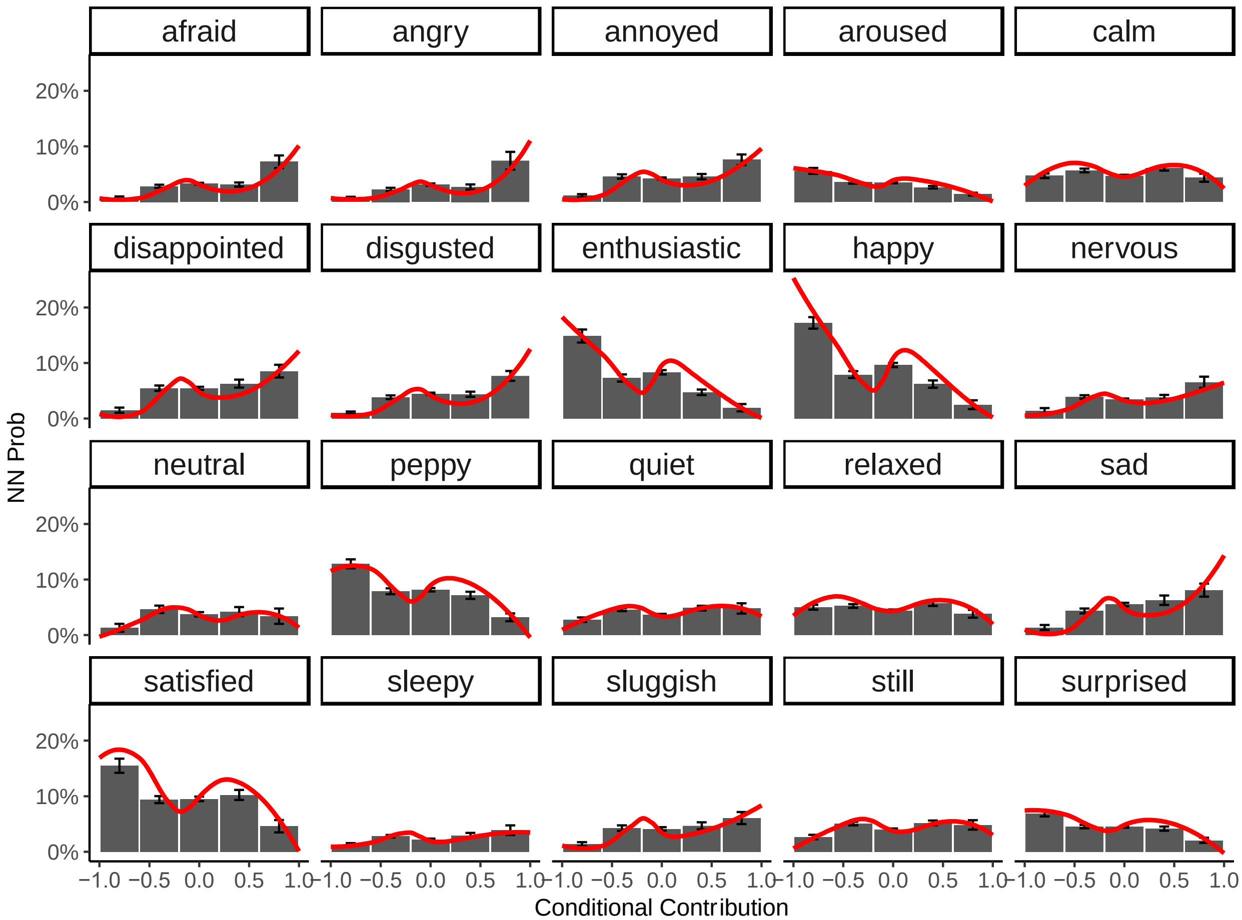
Supplementary Figure 11. Classifications by continuous contributions in the PD. *The neural network model trained on the emotion classification data was applied to the unlabeled emotion ratings from the Prisoner's Dilemma. Each data point was assigned a probability of each emotion class and is associated with a specific continuous contribution. Red lines reflect the fixed-effect predictions from separate mixed-effects regressions for each emotion class.*



Supplementary Figure 12. Relationship between contributions and emotion class probability. *Separate mixed-effects regressions were run for each emotion class predicting the probability of that emotion class as a function of contributions in the PD for participants (N = 306). Bars reflect the fixed-effect slope of contributions for each emotion class. Higher estimates indicate that increasing contributions increased the probability of this emotion class. Data are presented as beta coefficients while error bars reflect standard errors.*

3.3 Conditional Contribution Analysis

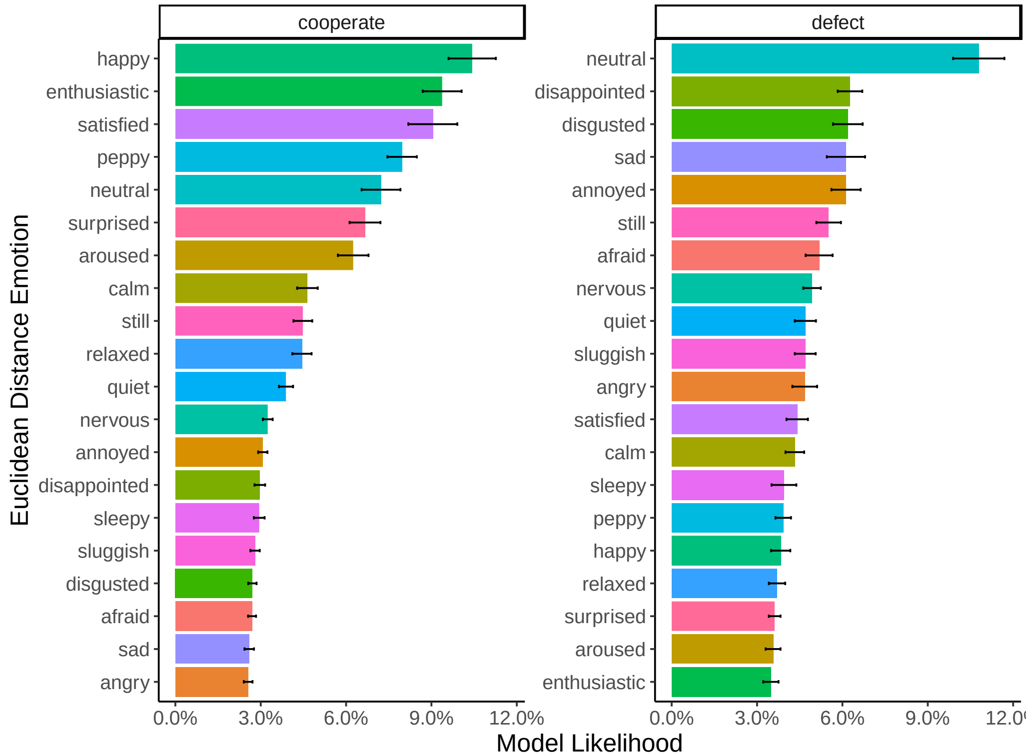
We also investigated how the probability of any specific emotion classification changed if we model conditional cooperation. Participants played a sequential Prisoner's Dilemma, meaning that participants contributions were made after their partners and therefore represent conditional cooperation. Because the relationship between conditional cooperation and neural network emotion probabilities is largely non-linear, we used separate loess regressions for each emotion class (Supplementary Figure 13). Results show that emotions have different relationships with conditional cooperation – for example, participants are very likely to feel happy when completely defecting against their partners (conditional contribution = -1) while emotions such as angry are most high when conditional cooperation is high (conditional contribution = +1). Although participants most often matched their partner's contribution, one explanation for these puzzling results is that some participants consistently contributed nothing $0 or everything $1, effectively deciding before seeing their partner's contribution, and these relationships may reflect post-decision emotions rather than pre-decision emotions.



Supplementary Figure 13. NN Probabilities by Conditional Cooperation in the PD. *Conditional cooperation is the difference between the participant's contribution and their partner. Conditional cooperation has been binned into five bars to visualize the non-linear relationship between conditional cooperation and neural network probabilities. Data are presented as mean values while error bars reflect 95% CIs. Red lines are loess regression lines using the continuous conditional cooperation measure for each emotion category for participants (N = 306).*

3.4 Euclidean Distance Analysis

We also developed an emotion classifier that is unique to each participant by calculating the Euclidean distance between all 20 emotion labels and the affective experience during each trial of the PD. Results show that neutral, disappointment, and disgust are the top three emotions associated with defecting (Supplementary Figure 14).

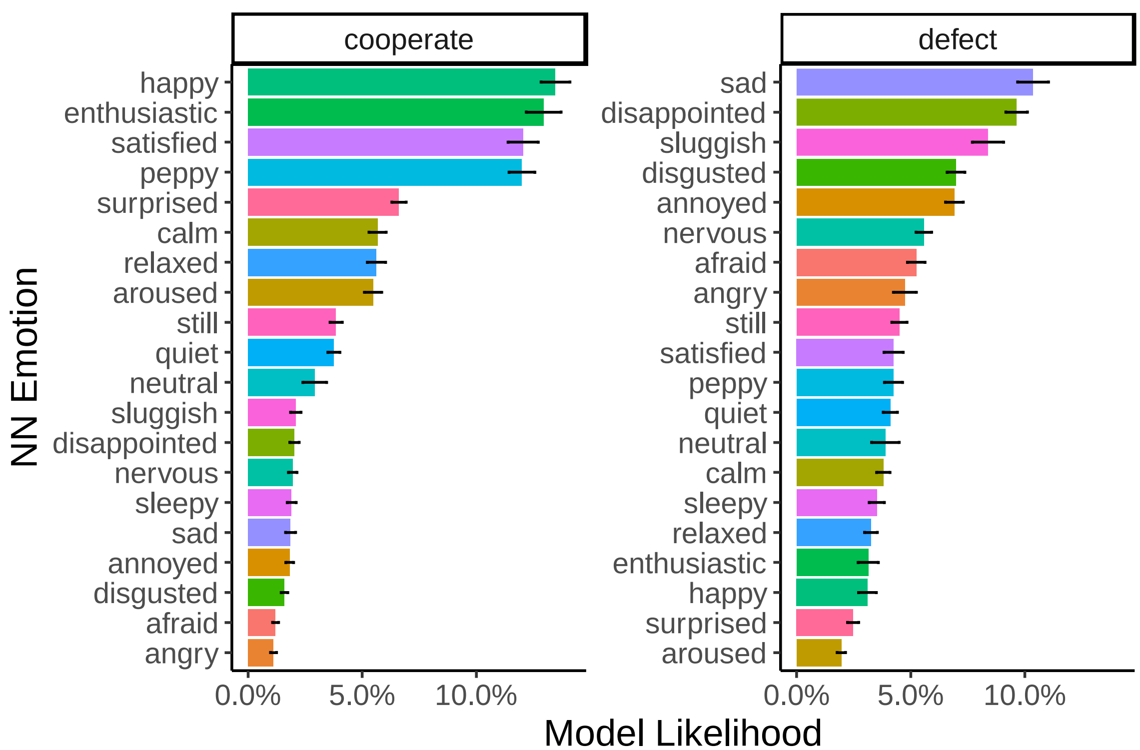


Supplementary Figure 14. Euclidean distance classifier for the PD*. Inverse distance weighting was used to calculate the probability of each emotion class for all unlabeled emotion ratings in the Prisoner's Dilemma. This classification was done separately for each participant (N = 306), and model likelihoods were averaged within participant and then across participants. Data are presented as mean values while error bars reflect 95% CIs.*

4 Experiment 3: Public Goods Game

4.1 Discrete Contribution Analysis

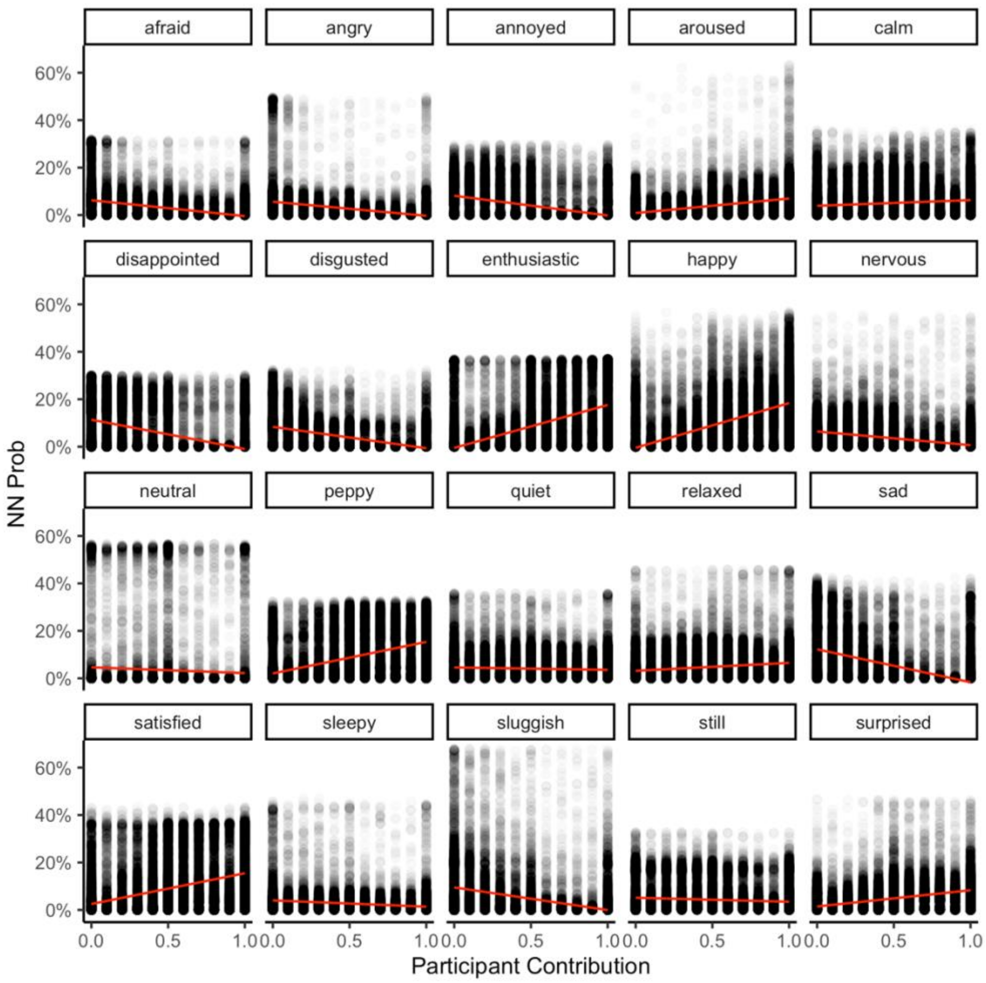
We binned continuous contributions ($0 - $1) made in the Prisoner's Dilemma (PD) into decisions to defect ($0 - $0.49) and cooperate ($0.50 - $1), and classified emotions experienced in the PD using the trained neural network from the emotion classification task. Results reveal that the top three emotions associated with decisions to defect are sadness (10.36%), disappointment (9.63%), and sluggishness (8.38%), with anger (4.74%) being identified as the 8th most likely emotion to be experienced in the Public Goods Game (Supplementary Figure 15). Paired t-tests revealed that sadness (t(443) = 12.9, $p < .001$, $d = .61$), disappointment ($t(443) = 14.0$, $p < .001$, $d = .67$), and sluggishness ($t(443) = 7.60$, $p < .001$, $d = .36$) were all significantly more likely to be associated with punishment than anger. When cooperating, the top three emotions most likely to be experienced were happiness, enthusiasm, and satisfaction.



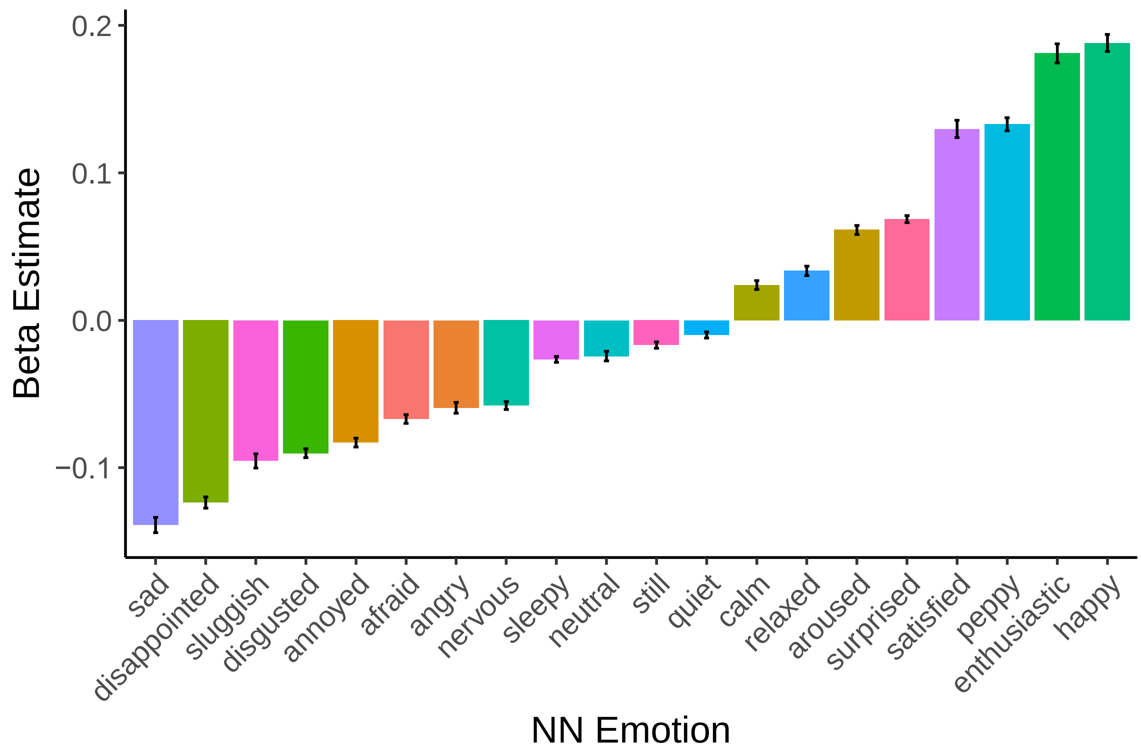
Supplementary Figure 15. Classifications by decisions to cooperate and defect in the PGG. *The neural network model trained on the emotion classification data was applied to the unlabeled emotion ratings from the Public Goods Game. Each data point was assigned a probability of each emotion class and model likelihoods were averaged within participant (N = 470) and then across participants. Data are presented as mean values while error bars reflect 95% CIs.*

4.2 Continuous Contribution Analysis

We again investigated how the probability of any specific emotion classification might change with continuous contributions ($0 - $1) in the PGG. We used separate mixed-effects regressions to examine the relationship between emotion probabilities and participant contributions. Results revealed that sadness, disappointment, and sluggishness are more sensitive to changes in a participant's contributions, when compared to anger (Supplementary Figure 16; Supplementary Figure 17).



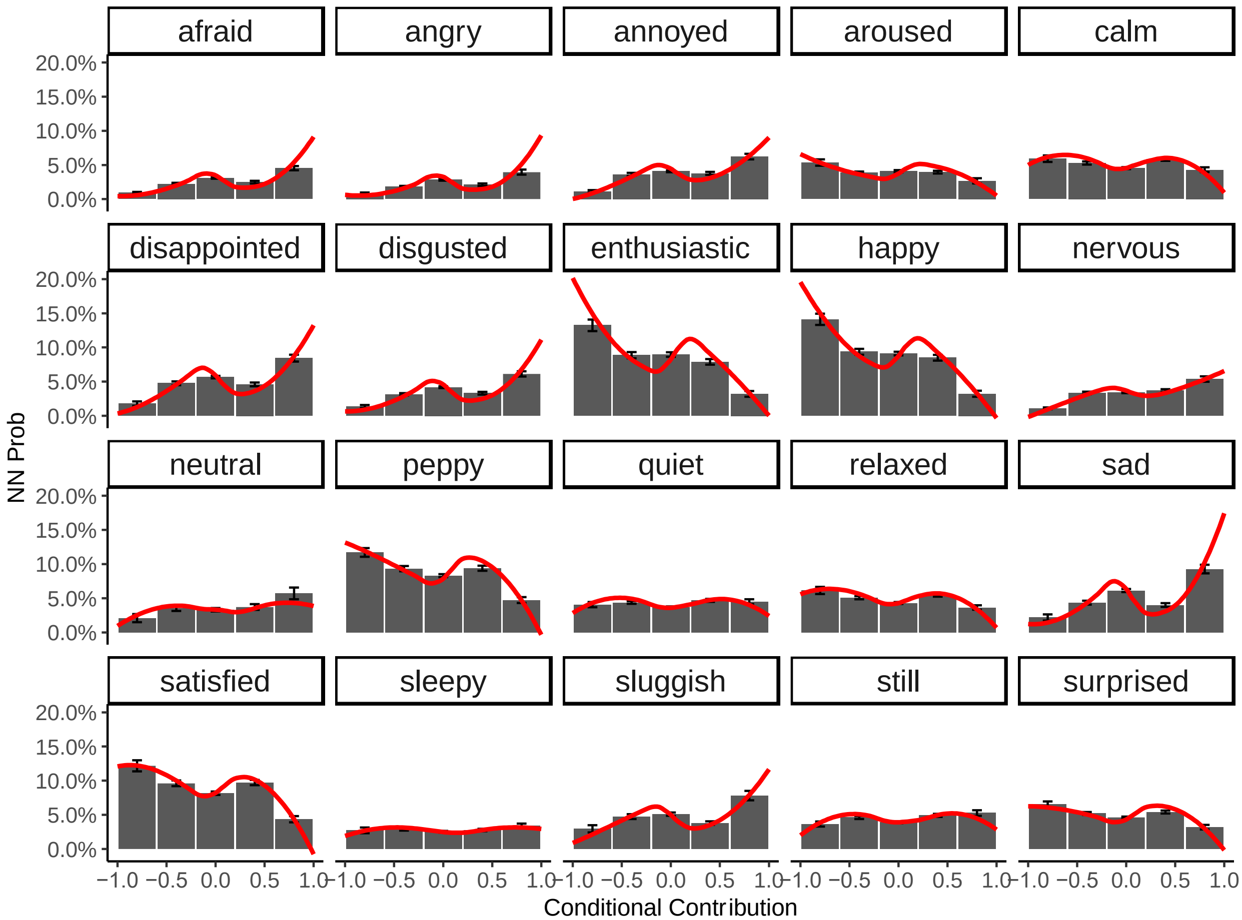
Supplementary Figure 16. Classifications by continuous contributions in the PGG. *The neural network model trained on the emotion classification data was applied to the unlabeled emotion ratings from the Public Goods Game. Each data point was assigned a probability of each emotion class and is associated with a specific continuous contribution. Red lines reflect the fixed-effect predictions from separate mixed-effects regressions for each emotion class.*



Supplementary Figure 17. Relationship between contributions and emotion class probability. *Separate mixed-effects regressions were run for each emotion class predicting the probability of that emotion class as a function of contributions in the PGG for participants (N = 470). Bars reflect the fixed-effect slope of contributions for each emotion class. Higher estimates indicate that increasing contributions increased the probability of this emotion class. Data are presented as beta coefficients while error bars reflect standard errors.*

4.3 Conditional Contribution Analysis

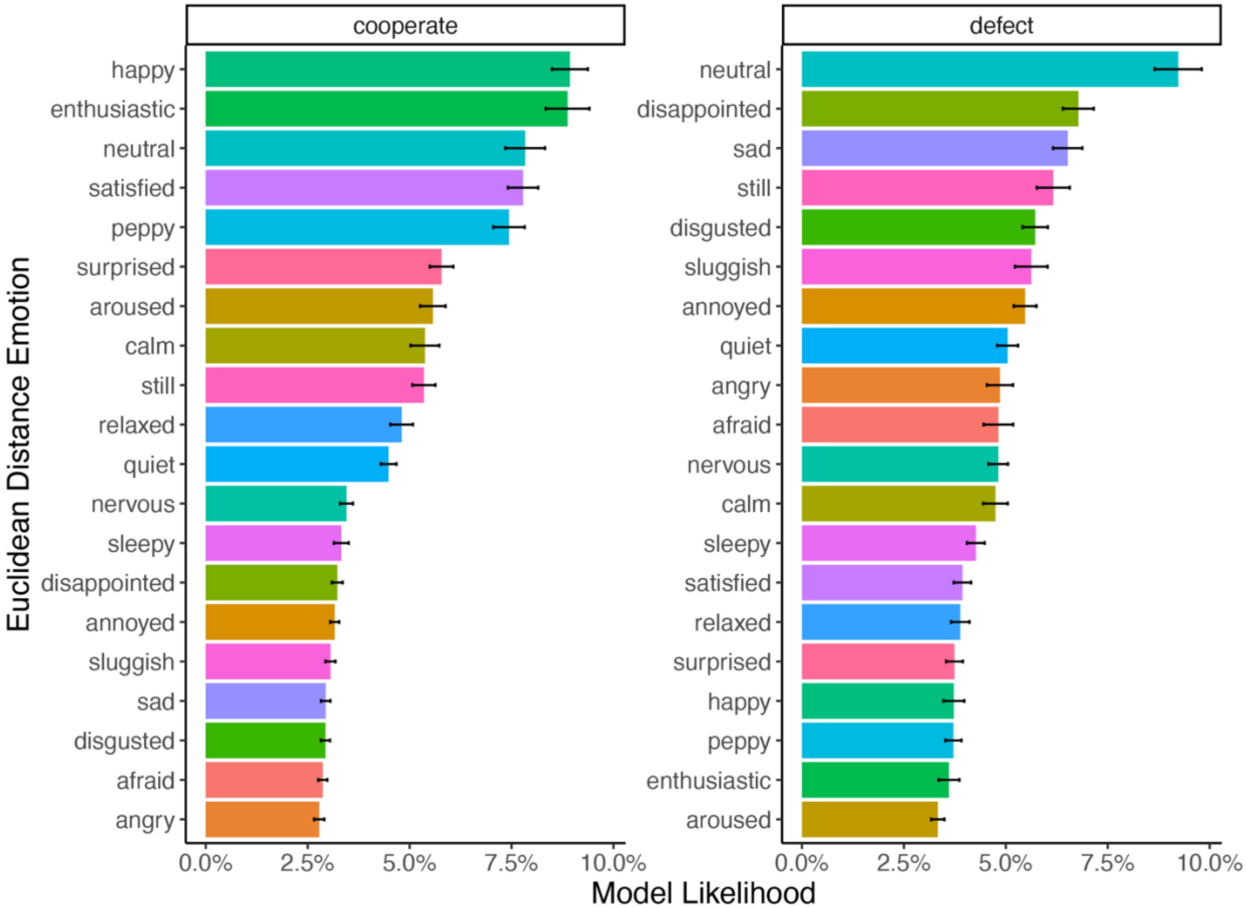
We also investigated how the probability of any specific emotion classification changed if we model conditional cooperation. Participants played a sequential Public Goods Game with three partners, meaning that participants contributions were made after their partners and therefore represent conditional cooperation. Participants were only shown the aggregate total of their partner's contribution, so we calculated conditional cooperation as the participant's contribution minus the average partner contribution (i.e., total contribution / 3). Because the relationship between conditional cooperation and neural network emotion probabilities is largely non-linear, we used separate loess regressions for each emotion class (Supplementary Figure 18).



Supplementary Figure 18. NN Probabilities by Conditional Cooperation in the PGG. *Conditional cooperation is the difference between the participant's contribution and the average partner's contribution (total contribution / 3). Conditional cooperation has been binned into five bars to visualize the non-linear relationship between conditional cooperation and neural network probabilities. Data are presented as mean values while error bars reflect 95% CIs. Red lines are loess regression lines using the continuous conditional cooperation measure for each emotion category for participants (N = 470).*

4.4 Euclidean Distance Analysis

We also developed an emotion classifier that is unique to each participant by calculating the Euclidean distance between all 20 emotion labels and the affective experience during each trial of the PGG. Results show that neutral, disappointment, and sadness are the top three emotions associated with defecting (Supplementary Figure 19).



Supplementary Figure 19. Euclidean distance classifier for the PGG. *Inverse distance weighting was used to calculate the probability of each emotion class for all unlabeled emotion ratings in the Public Goods Game. This classification was done separately for each participant (N = 470), and model likelihoods were averaged within participant and then across participants. Data are presented as mean values while error bars reflect 95% CIs.*

Chapter 3

Emotional responses to prosocial messages increase willingness to self-isolate during the COVID-19 pandemic

Joseph Heffner, Marc-Lluís Vives & Oriel FeldmanHall
Published in Personality and Individual Differences (2021)

Introduction

In the span of just a few months, COVID-19 has ripped through almost every country, infecting 30 million people, killing close to a million individuals as of September 17th, 2020 (John Hopkins University, 2020), and severely crippling dozens of economies. Without a vaccine in hand, it seems that the virus can only be slowed by extreme behavioral change and societal coordination (Arenas et al., 2020). Some countries, like South Korea, were quick to respond by instituting enforced quarantines and entreating citizens to practice social distancing (Beech, 2020; Fisher & Sang-Hun, 2020). Other countries, like the United States and the United Kingdom, were reluctant to impose widespread shelter-in-place measures (The Associated Press, 2020). In the United States, for example, individual states began gradually issuing social isolation practices to combat the spread of the virus through the months of March and April of 2020 (Mervosh, Lu, & Swales, 2020). In both cases, the countries hoped their citizens would readily comply with public health messages. Preliminary reports, however, show vast differences in people's willingness to practice measures that can reduce pathogen transmission (Lunn et al., 2020).

At present, public health advisors, such as the World Health Organization, argue that mitigating the spread of COVID-19 necessitates people swiftly adapt and change their usual habits to obey new social distancing measures (World Health Organization, 2020). Problematically, social distancing measures increase unemployment rates (Coibion, Gorodnichenko, & Weber, 2020), influence work productivity, and acutely affect mental wellbeing (Kawohl & Nordt, 2020). Thus, the actions needed to reduce the spread of COVID-19 are in direct opposition to functioning daily life. This poses a critical challenge for accomplishing extreme behavior change compliance, especially in such large populations.

Decades of research show that emotional engagement is a critical component of behavior change (Bagozzi & Pieters, 1998; Cooper & Nisbet, 2016; Hartley & Phelps, 2010; Nabi, 2007; Perugini & Bagozzi, 2001), which is why it is often employed in public health campaigns (Dillard & Nabi, 2006; Lang & Yegiyan, 2008; Nabi, 1999, 2002; Zeelenberg & Pieters, 2006). However, the relationship between emotion and behavior change is not straightforward (O'Keefe, 2012). For instance, tailoring messages to evoke a specific emotional response can backfire: When public service announcements about binge drinking evoke shame—rather than the intended guilt—vulnerable populations can increase their alcohol consumption (Duhachek, Agrawal, & Han, 2012). Fear is also notoriously fickle in creating successful behavior change (Hastings, Stead, & Webb, 2004; Leventhal, 1970; Petty & Cacioppo, 1996). Some research shows that only those most at risk for certain behaviors, such as drunk driving, are least responsive to messages with fearful language (Tay & Ozanne, 2002). While widespread and rapid adoption of preventative measures is unlikely to occur without messages that include emotional appeals (Myers, Nisbet, Maibach, & Leiserowitz, 2012), it is crucial that current public health officials and researchers understand the relationship between emotional engagement and different persuasive messages related to COVID-19.

Despite the complexity of the relationship between emotion and behavior, some media outlets have been leveraging fear language in order to motivate people to stay home and socially distance. A recent article, for example, highlighted grim outcomes, staggering death tolls, and an inability for an overwhelmed health system to treat citizens (Pueyo, 2020). There is good reason to specifically focus on fear related to COVID-19 (Feldman & Hart, 2015; Moser, 2010; Nisbet, 2009): Evoking fear can potently effect attitudes and behaviors (Tannenbaum et al., 2015), likely because fear can enhance attention towards the message (Baron, 1994) and increase perceptions of threat (Leiserowitz, 2006). However, the relationship between fear and disease prevention behaviors is not straightforward (Hastings, Stead, & Webb, 2004). A message that is perceived as too threating can cause people to engage in defensive avoidance, which leads them to disregard the message altogether (Janis & Feshbach, 1953). Indeed, across a host of behaviors, a message that evokes too much (Janis & Feshbach, 1953; Krisher, Darley, & Darley, 1973), too little (Boster & Mongeau, 1984; Witte & Allen, 2000), or in some cases, any amount of fear at all (O'Neill & Nicholson-Cole, 2009), can fail to produce any noticeable behavioral change. Thus, while the use of fearful language is widely adopted as a means for behavior change, the evidence to date illustrates that its efficacy is variable (O'Keefe, 2012).

On the other hand, appeals that use prosocial rather than threatening language can play a potent role in the efficacy of public health campaigns (Lewis, Watson, White, & Tay, 2007), serving as a distinct contrast to fear-based appeals. For example, describing prosocial actions that can lead to positive outcomes in the face of public health problems can produce positive emotions, such as hope or joy (Nabi, Gustafson, & Jensen, 2018; Ojala, 2012), which can increase reception to the message by reframing the issue as being more personally relevant (Monahan, 1995). Indeed, some recent research illustrates that prosocial public health messages that underscore behaviors linked to societal and communal benefits (e.g., help protect your fellow citizens)—rather than focusing on behaviors that only benefit the self (e.g., protect yourself)—may be an especially effective method (Kelly & Hornik, 2016; Li, Taylor, Atkins, Chapman, & Galvani, 2016) for communicating public health recommendations related to COVID-19 (Jordan, Yoeli, & Rand, 2020).

One additional difficulty in designing public health messages is the heterogeneity of emotional responses to interventions (Carey & Sarma, 2016). Although typically outside the scope of public health research, characterizing how stable personality traits interact with emotional experiences can provide inroads for understanding the link between emotions and a message's efficacy. For example, the biological theory of personality explores how extraversion and neuroticism are linked with the body's physiological response (Eysenck & Eysenck, 1991), and these traits generally correlate with positive and negative mood, respectively (Costa & McCrae, 1980; Rusting & Larsen, 1997). Whereas neurotic tendencies are linked to increased emotional arousal (Haas, Constable, & Canli, 2008; Kehoe, Toomey, Balsters, & Bokde, 2012), extraverts are less likely to be as reactive to arousing stimuli. When considered within the framework of public health messaging, this suggests that neuroticism, and not extroversion, may predict stronger arousal responses to emotionally evoking messages.

Presently, it is unknown whether messages using threating or prosocial language are equally effective in promoting changes in willingness to self-isolate regarding COVID-19. Research on public health campaigns typically examine specific emotions (Nabi et al., 2018; Ojala, 2012; Tannenbaum et al., 2015; Witte & Allen, 2000), but these approaches constrain a person's emotional experiences by limiting them to identifying with a set of discrete emotions pre-selected by the researcher. For example, asking how afraid one feels after reading a message imposes an emotional structure that the participant “ought” to feel afraid. Scaffolding the question in such a manner assumes that these emotional words are interpreted in similar ways across individuals, and may even influence how the very emotion is experienced (Kassam & Mendes, 2013). Here, we circumvent these issues by using a model of emotion that avoids specific emotion states and partitions emotional experiences into a two dimensional space: the affective dimensions of valence (pleasurableness) and arousal (alertness/activation; Russell & Barrett, 1999). Using this approach, we can characterize the heterogeneity in emotional responses to both threat and prosocial appeals related to COVID-19, and directly relate emotional engagement on the independent dimensions of valence and arousal to message efficacy. Given that past research suggests that the intensity of emotional engagement (i.e., arousal) increases learning, memory, and attention (Kensinger & Corkin, 2004; Reisberg & Heuer, 1992; Storbeck & Clore, 2008), we posited that increases in emotional responses would result in greater compliance. However, because of the inconsistent relationship between evoked fear and behavioral change in prior research, we were agnostic as to whether stronger valence and arousal reactions to the threat intervention, compared to the prosocial intervention, would result in more willingness to self-isolate regarding COVID-19.

Methods

Participants

On March 24th, 2020, we began recruitment through the online site Prolific to collect a representative United States sample (based on sex, age, and ethnicity; Prolific Team, 2019) of N = 1000. Because effect sizes of persuasion on behavior are highly variable, our study used a conservative estimate of the smallest effect size of interest (Lakens, 2017). Using a lower equivalence bound of d = -0.10 and upper bound of d

= 0.10, our study was well powered (87%) to detect effect sizes with a greater absolute magnitude than 0.1 with an alpha of 0.05. Participants received monetary compensation and provided informed consent in a manner approved by Brown University's Institutional Review Board. The experiment was conducted within a week of the COVID-19 infection reports in the United States reaching 10,000 (John Hopkins University, 2020). We only recruited U.S. participants to ensure that national messages and questionnaires specific to the United States would be relevant. For example, on March 13th, the White House released a proclamation declaring a national state of emergency related to the COVID-19 outbreak (The White House, 2020) and on March 16th, social distancing guidelines were issued in the United States (The White House & Centers for Disease Control and Prevention, 2020). Using the preregistered exclusion criterion that aimed to ensure high quality data, we excluded 45 individuals' data using a conservative measure of noncompliance based on instructions for an emotion classification task (see *Measuring Emotional Experiences* for a description of the task). This resulted in a final sample of 955 participants recruited between March 24th and March 26th, 2020 (506 females; age M = 44.8, SD = 15.9). Participants reported being 73.0% White, 13.4% Black, 4.4% East Asian, 3.9% Hispanic / Latinx, 2.1% South Asian, 1.6% Mixed Race, 0.4% Native American, 0.3% Middle Eastern, and 0.9% Other.

General procedure

Here we detail every measure that participants responded to, however, only the intervention measures, personality measures (BFI-2-S; Soto & John, 2017) and a questionnaire on COVID-19 preventative behaviors (e.g., "I stayed at home", which provided a baseline for COVID-19 self-isolation behavior), were analyzed for this experiment (all detailed below). All other measures were collected for another experiment, whose hypotheses and methods were preregistered on OSF (https://osf.io/y2uj6). All participants completed a series of tasks and questionnaires in the following order: an emotion classification task, a variety of self-report questionnaires with a randomly presented order including the emotion regulation questionnaire (Gross & John, 2003), interpersonal regulation questionnaire (Williams, Morelli, Ong, & Zaki, 2018), extraversion and neuroticism subscales of the Big Five Inventory-2-S (Soto & John, 2017), intolerance of uncertainty (Carleton, Norton, & Asmundson, 2007), and clinical measures of depression (Radloff, 1977), anxiety (Spitzer, Kroenke, Williams, & Lowe, 2006), and alexithymia (Bagby, Parker, & Taylor, 1994), a questionnaire that assessed their knowledge of COVID-19, a fear intervention, questionnaires that assessed behavioral responses towards COVID-19, fear of COVID-19, media consumption of COVID-19, motives related to COVID-19 behaviors, social support related to COVID-19, information about work related to COVID-19, an altruism intervention, and demographics.

Interventions

In a within-subject design, participants were given two prompts we created in the following order: The threat intervention followed by the prosocial intervention. The threat intervention was inspired by a recent Medium article (Pueyo, 2020) that tapped into fear of COVID-19: *"The coronavirus is coming for you. When it does, your healthcare system will be overwhelmed. Your fellow citizens will be turned away at the hospital doors. Exhausted healthcare workers will break down. Millions will die. The only way to prevent this crisis is social distancing today."* This generated threat message contains two traditional components of threat appeals that emphasize: 1) the severity of the issue through negative consequences, and 2) the likelihood these consequences will occur to the reader (Dillard, Li, Meczkowski, Yang, & Shen, 2016).

After reading this prompt, participants were asked three questions: (a) How does this statement make you feel? (responses recorded using a granular emotion measure, see details below); (b) On a scale from 0 (not at all) to 100 (completely), how willing are you to self-isolate?; (c) On a scale from 0 (no change) to 100 (a lot of change), how much does the previous statement change your willingness to self-isolate? In the second prompt (which was temporally spaced by multiple questionnaires in between), participants were given a prosocial intervention that was designed to be as similar as possible in structure to the threat prompt, but which emphasized prosocial actions: *"Help save our most vulnerable. Together, we can stop* *the coronavirus. Everyone's actions count, every single person can help to slow the crisis. We have the tools to solve this problem. Together, by self-isolating we can save millions of lives."* This prosocial message focused on internal efficacy (how the individual can take successful action to mitigate the spread of COVID-19) and response efficacy (emphasizing the effectiveness of the group working together; Hart & Feldman, 2014). After this prompt, participants were again asked the three questions denoted above.

Measuring emotional experiences

After reading both the threat and prosocial appeals, participants reported their affective experiences using the *dynamic Affective Representation Mapping* (dARM) tool (a measure we have used in our work; (Heffner, Son, & FeldmanHall, under revision)), which was adapted from the affect grid used in past research (Russell, Weiss, & Mendelsohn, 1989). This measure allows participants to rate their affective experiences on a subjective map where the horizontal axis characterizes an unpleasant-pleasant dimension (i.e., valence), and the vertical axis characterizes a low-high activation dimension (i.e., arousal). The dARM has a sampling resolution of 500 x 500 pixels, enabling us to measure fine-grained self-reports of both the valence and arousal dimensions without forcing discrete emotional labels on their experiences. To ensure participants were able to effectively use the dARM to rate their emotional experiences after appeals ("How does this statement make you feel?"), participants completed an emotion classification task at the beginning of the experiment. The emotional classification task asked participants to rate 20 canonical emotion words (e.g., angry, sad, happy) using the dARM measurement. Critically, participants were told where to rate a specific feeling, neutral, in the instructions as an attention check: "*The center of the square represents a neutral, average, everyday feeling. It is neither positive nor negative*". Our preregistered exclusion criterion was to remove participants who failed to rate neutral within a 100 x 100 pixel square around the center (N = 45/1000).

Analysis

We used linear mixed-effects regressions to predict participants' self-reported 1) willingness to self-isolate, and 2) change in willingness to self-isolate after reading the interventions. Predictor variables were participant's emotional ratings on the dARM, separated by the arousal and valence dimensions, as well as the type of intervention (threat/prosocial). Separate regressions were run for predicting willingness to self-isolate (labeled "willingness") and change in willingness (labeled "change"). For the personality analyses, we used separate linear mixed-effects regressions to predict participants' arousal and valence ratings as a function of the personality trait and intervention type. All regressions were run using the nlme package in R (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2020).

Results

To examine the effectiveness of the fear and prosocial interventions, we first examined current reported self-isolation behavior. We found that, on average, people reported staying at home 87.3% of the time because of COVID-19 ("I stayed home" ranging from 0 – not at all to 100 – all the time). Although people's initial willingness was heavily skewed (skewness = -2.80) such that most people were already reporting engaging in self-isolationist measures, we can still examine whether the two interventions encouraged people to engage even more in self-isolation.

Prosocial and threat interventions are equally effective in increasing willingness to self-isolate

To create a measure of each intervention's effectiveness, we subtracted reports of current self-isolation from reported willingness to self-isolate after reading the threat and prosocial interventions (both scales ranged from 0-100). Comparing intervention' scores to 0 (i.e., no effect of intervention) revealed that both the threat intervention (M = 6.44, SD = 15.41; $t(954)$ = 12.92, $p < .001$; Cohen's d = .42) and prosocial intervention (M = 6.50, SD = 15.71; $t(954)$ = 12.79, $p < .001$; Cohen's d = .41) increased willingness to self-isolate, confirming the efficacy of both interventions. Importantly, these results remain the same when we use an aggregate measure of all preventative COVID-19 behaviors, rather than a single item assessing willingness to stay home (threat Cohen's d = .31, prosocial Cohen's d = .30). We then examined whether the threat or prosocial intervention produced differences in people's reported willingness to self-isolate (termed "willingness"; Fig. 1A), as well as their reported change in self-isolation behavior after reading the intervention (termed "change"). Although participants reported high levels of willingness to self-isolate after reading both the threat intervention (M = 93.75, SD = 12.96) and the prosocial intervention (M = 93.81, SD = 13.43), a paired sample t-test showed the two interventions did not produce significantly different reports of willingness ($t(954) = 0.25$, p = .81) or changes in self-isolation ($t(954) = 0.17$, p = .87). The correlation between willingness and change was significant for both the prosocial ($r(953) = -0.07$, p = .04) and threat interventions ($r(953) = -0.07$, p = .04). Together, these results illustrate that both prosocial and threat interventions nudged willingness to self-isolate in comparable ways to help mitigate the spread of the virus.

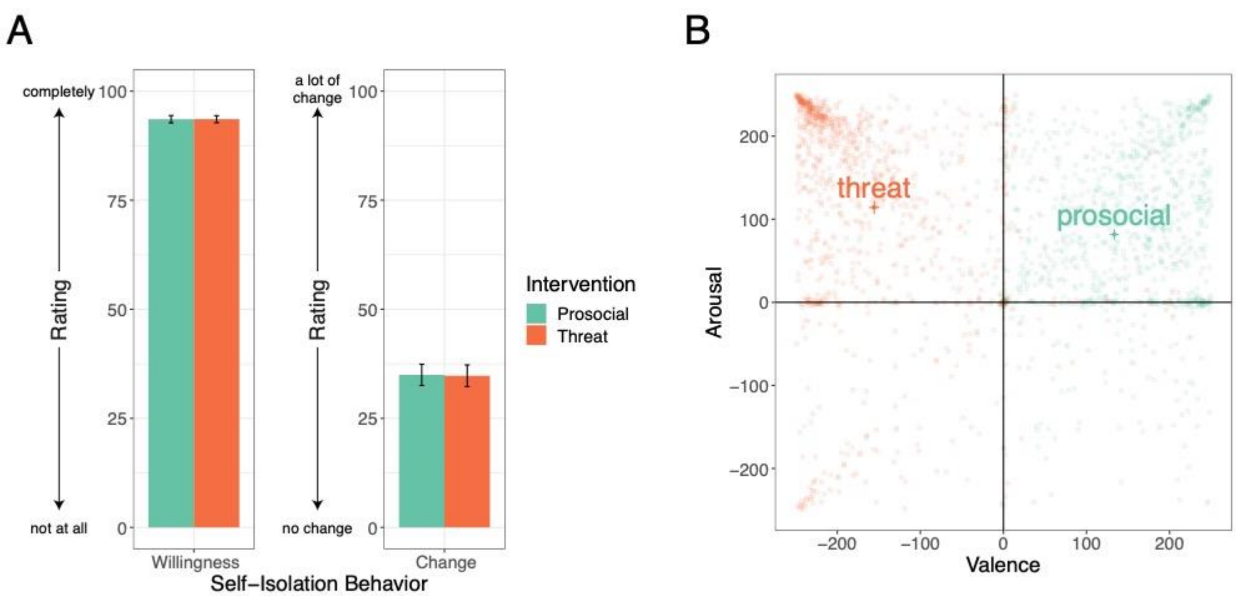


Figure 1. Intervention Results. A) Self-isolation behavior after reading each intervention*. Participants reported how willing they were to self-isolate from 0 (not at all) to 100 (completely) and how the intervention changed their willingness to self-isolate from 0 (no change) to 100 (a lot of change). Participants report similar levels of willingness and change in self-isolation to the prosocial and threat interventions.* ***B). Emotional experiences after interventions****. Participants reported how each*

intervention made them feel on the dynamic Affective Representation Mapping (dARM) measure, which simultaneously captures experienced valence and arousal at a granular level. Raw data has been plotted as transparent dots and the group averages are plotted below the intervention labels. All error bars are 95% confidence intervals (CIs).

Threat interventions evoke stronger valence and arousal responses compared to prosocial interventions

While the interventions were similarly effective (Fig. 1A), examining the emotional reactions to both interventions revealed they were associated with distinct emotional experiences. The average emotional response to the threat intervention was very unpleasant and highly arousing while the average emotional response to the prosocial intervention was fairly pleasant and moderately arousing (Fig. 1B). The emotional responses to the threat intervention were heavily clustered in the upper-left corner of the dARM, indicating more homogeneity in emotional responses compared to the prosocial intervention responses. Formal tests comparing experienced arousal and valence between the two interventions revealed that the threat intervention was experienced as significantly more arousing (M = 115.46, SD = 126.60) than the prosocial intervention (M = 81.88, SD = 99.37; paired sample $t(954) = 7.90$, $p < .001$; Cohen's d = .26). Moreover, while the threat intervention was unsurprisingly significantly more negatively valenced (M = -158.67, SD = 94.28) than the prosocial intervention (M = 134.60, SD = 90.95; $t(954) = -68.50$, $p < .001$; Cohen's d = 2.22), it was critically experienced as more unpleasant than the prosocial intervention was experienced as pleasant (absolute value of valence, $t(954) = 7.56$, $p < .001$; Cohen's d = .24). This suggests that participants had a stronger emotional reaction on both dimensions to the threat intervention than the prosocial intervention.

Extraversion and neuroticism explain emotional responses to prosocial and threat interventions, respectively

Given the heterogeneity of emotional responses to the two interventions (Fig. 1B), we next examined whether personality traits known to predict positive and negative moods—extraversion and neuroticism, respectively (Costa & McCrae, 1980; Rusting & Larsen, 1997)—explain the observed emotional variance. We found that neuroticism interacted with intervention type to predict increasing arousal (interaction $\beta = -0.10 \pm 0.03$, $p < .001$) but not valence (interaction $\beta = -0.02 \pm 0.02$, $p = .52$), and this was uniquely carried by the threat intervention ($\beta = 0.07 \pm 0.02$, $p = .004$; prosocial intervention: $\beta = -0.03 \pm 0.02$, $p = .19$). There was also an observed main effect of neuroticism predicting negative valence for both intervention types (threat $\beta = -0.05 \pm 0.02$, $p = .006$; prosocial $\beta = -0.06 \pm 0.02$, $p < .001$), suggesting that individuals higher in neuroticism generally experienced more negative emotions to the interventions. Although less predictive, extraversion also interacted with intervention type to predict increasing arousal in the opposite direction as neuroticism (interaction $\beta = 0.06 \pm 0.03$, $p = .02$), and this was uniquely driven by the prosocial intervention ($\beta = 0.05 \pm 0.02$, $p = .04$) and not the threat intervention ($\beta = -0.01 \pm 0.02$, $p = .56$). Finally, we observed that greater extraversion predicted increasingly positive valence for the prosocial intervention ($\beta = 0.06 \pm 0.02$, $p = .001$) but not the threat intervention ($\beta = 0.02 \pm 0.02$, $p = .31$). However, because of a nonsignificant interaction between extraversion and intervention type predicting valence (interaction $\beta = 0.04 \pm 0.02$, $p = .10$), these simple effects should be interpreted with caution.

The efficacy of prosocial—but not threat—interventions depend on degree of emotional engagement

Examining how these emotional responses influenced willingness to self-isolate revealed that the strength of experienced arousal and valence was more associated with willingness to self-isolate for the prosocial intervention compared to the threat intervention (arousal: interaction $\beta = 0.07 \pm 0.03$, $p = .023$; valence: interaction $\beta = 0.16 \pm 0.05$, $p < .001$; Fig. 2). Indeed, the fact that the simple effects of the threat intervention (dark red and dark purple lines in Fig. 2) were not significant suggests that the efficacy of the threat intervention does not rely on the strength of the emotional response, whereas the prosocial intervention does. A similar behavioral pattern was found for changes in self-isolation (Fig. 3), where the effect of valence on behavior change was significantly higher for the prosocial intervention than the threat intervention (interaction $\beta = 0.24 \pm 0.06$, $p < .001$). However, unlike before, the relationship between arousal and change was similar across both interventions (interaction $\beta = 0.02 \pm 0.04$, $p = .644$), suggesting that increases in arousal for both interventions lead to more intention to change behavior.

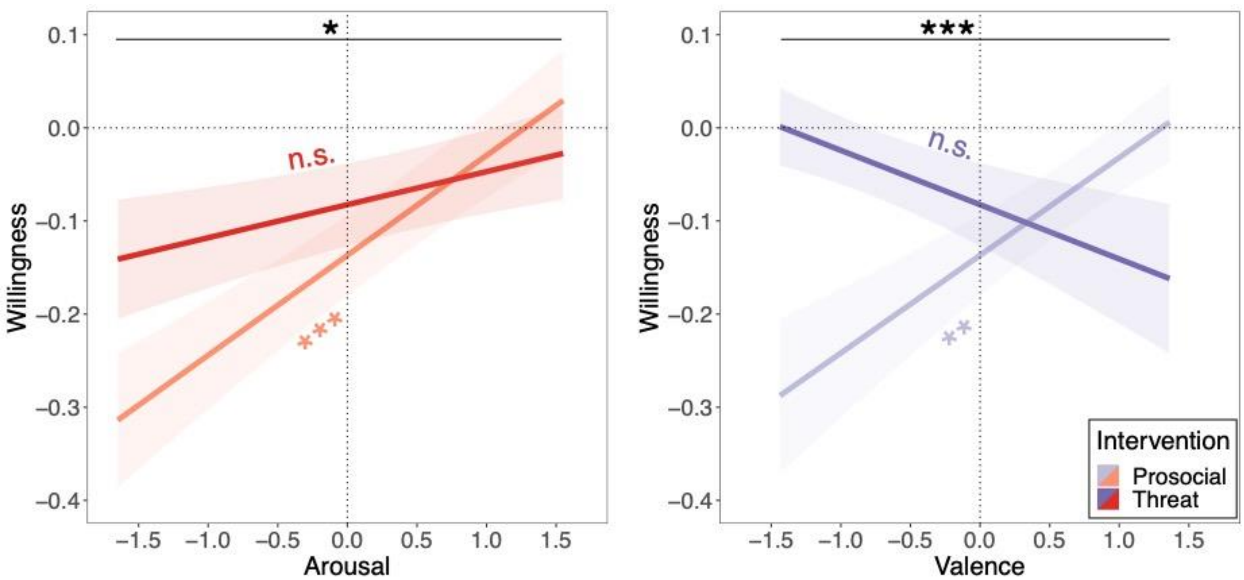


Figure 2. Emotional experience predicts reported willingness to self-isolate after prosocial intervention. *Willingness to self-isolate is plotted for arousal and valence after reading the prosocial and threat interventions. 'Willingness' has been normalized (standardized and mean-centered) while arousal and valence have been standardized without being mean-centered (as the 0 point on the scale reflects a neutral feeling). Lines represent regression fits and shaded areas reflect ±1 standard errors (SEs).*

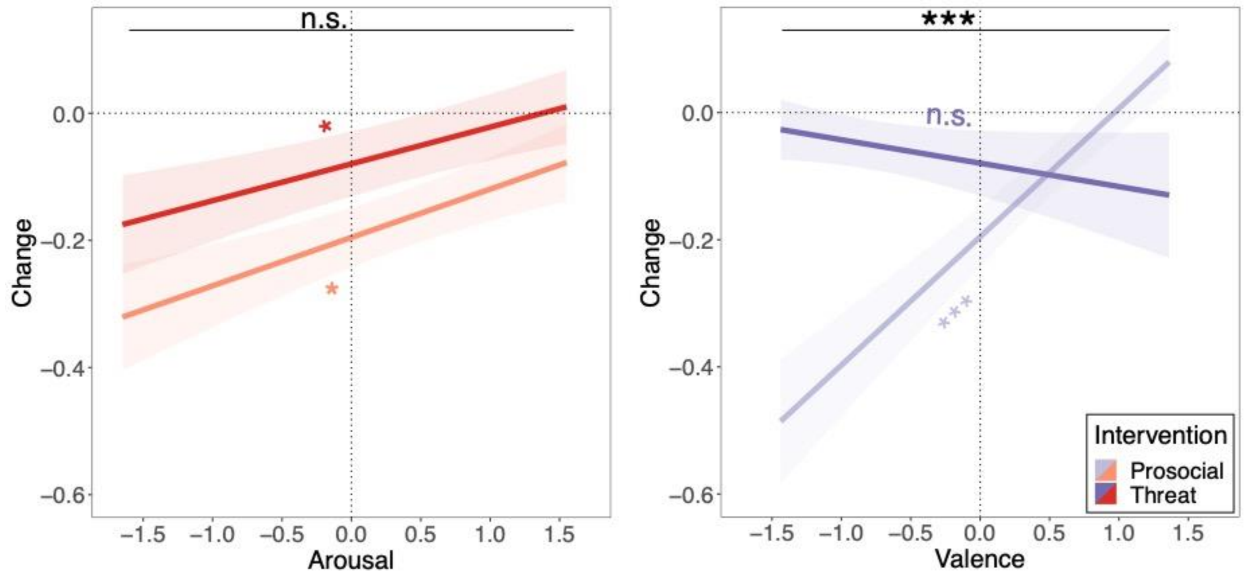


Figure 3. Emotional experience predicts reported changes in self-isolation after prosocial intervention. *Change in reported self-isolation is plotted for arousal and valence after reading the prosocial and threat interventions. Change has been normalized (standardized and mean-centered) while arousal and valence have been standardized without being mean-centered (as the 0 point on the scale reflects a neutral feeling). Lines represent regression fits and shaded areas reflect ±1 standard errors (SEs).*

Discussion

The effectiveness of public health messages is crucial for successfully combating large public health crises such as the COVID-19 pandemic. Problematically, the behaviors associated with preventing the spread of the virus are difficult to adhere to, as they include vigilant hand washing, donning facial masks, and most disruptively, practicing extreme social distancing measures. This makes it challenging for public health officials to create messages that are effective in motivating behavior change. Here, we explore how emotion shapes the efficacy of two different persuasive appeals, one that highlights threat and one that emphasizes prosociality. Unlike previous research that has found prosocial frames to be more effective than threat frames (Shen, 2011), we find that both threat and prosocial messages were equally effective in stimulating willingness to engage in disease prevention health behaviors. However, while threat messages created a stronger emotional reaction (which were more negative and arousing) than the prosocial message, the efficacy of the threat intervention depended less on the strength of the emotional response compared to the prosocial intervention. In contrast, the prosocial message was more effective at boosting willingness to self-isolate if it produced a strong, positive, and arousing emotional response.

These findings reveal that although threat and prosocial interventions were similarly successful in changing people's self-isolation intentions, they do not operate from the same emotional mechanisms. While successful prosocial messages depend on strong, positive emotional engagement, effective threat messages leveraging fear-mongering language are less reliant on the strength of emotional reactions. Given the lack of observable relationship between emotion and reported willingness to self-isolate in response to fear-mongering language, other mechanisms such as a negativity bias (Rozin & Royzman, 2001), or selective attention to negative information (Carretié, Mercado, Tapia, & Hinojosa, 2001), may subserve the efficacy of fear messaging. Moreover, because stronger negative emotional responses did not yield influence on willingness to self-isolate, designing a message with more graphic and emotionally evocative language would likely not improve the success of a fear-mongering appeal, but it may increase the efficacy of prosocial messages. Since self-isolation and monetary hardship related to economic downturns can result in increases in depression and anxiety (Brooks et al., 2020), changing behavior without resorting to fear-mongering tactics would be important for public health officials to consider when designing public service announcements. Indeed, it is possible that messages associated with positive emotions may help buffer against any unnecessary increases in clinical mood disorders. Simply put, in a situation which may already exacerbate anxiety and depression, messages that promote behavioral change while simultaneously appealing to positive emotions are needed now more than ever.

However, the observed heterogeneity of emotional responses to these messages suggests a lingering challenge in fine-tuning the emotional language of either a threatening or prosocial message. Although exploratory in nature, we found evidence that two commonly measured personality traits, extraversion and neuroticism (Hamann & Canli, 2004) explain some of the emotional variance. Dovetailing with past research illustrating a link between positive and negative affect and extraversion and neuroticism, respectively (Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002), our results reveal that neuroticism uniquely mapped onto arousal for the threatening intervention while extraversion uniquely mapped onto arousal for the prosocial intervention. Furthermore, neuroticism and extraversion generally predicted negative and positive valence across the intervention types. Overall, this suggests that individuals high on neuroticism will respond more strongly to threat-based messages while extraverts will engage more with the prosocial ones. Although a clear limitation is that we did not examine the full spectrum of personality traits, these findings may help policy makers consider the type of public health message given the demographics of a specific population.

It is worth noting, however, that participants in our studies were simply asked to report their willingness to change their behaviors. Research on message interventions illustrates that reported behavior change does not always coincide with actual behavior changes in the real world (Gibbons, Gerrard, Ouellette, & Burzette, 1998). Although previous work has demonstrated that perceived message efficacy is a relatively good measure of actual effectiveness (Dillard, Shen, & Vail, 2007; Dillard, Weber, & Vail, 2007), it will be important to confirm that these results generalize to actual COVID-19 related behaviors, where readers are being bombarded with many different messages and likely exhibit divided attention when consuming news or reading public health messages. Furthermore, it is also critical to highlight that, while the rapid transmission of COVID-19 is unfolding on a global scale, our sample was limited, by design, to the United States. As there are known cultural differences in how emotion is conceptually represented (Jackson et al., 2019) and expressed (Gendron, Roberson, van der Vyver, & Barrett, 2014), caution should be taken when interpreting the widespread applicability of these results since findings may not translate cross-culturally. One additional limitation of the current design is that participants were given the threat message first, followed by the prosocial appeal. While future work should consider counterbalancing the messages, in general, fixed-order effects typically present minimal to no changes in results (Sauer, Auspurg, & Hinz, 2020).

As of the beginning of September 2020, the United States had still not achieved widespread compliance with social isolationist measures (Canipe, 2020; Fitzpatrick & DeSalvo, 2020), despite repeated calls for citizens to shelter in place from specific States. To help speed the global goal of “flattening the curve” (Qualls et al., 2017), governments and public health officials need to find the most effective messaging for stimulating behavioral compliance. While threat appeals to mobilize society during this time might be tempting to motivate behavioral compliance, we found that prosocial calls to action not only created more positive emotions, but they also elicited just as much willingness to self-isolate compared to deploying threatening language. Although these are preliminary results, it suggests that when collaborative efforts are needed to fight a global pandemic, interventions that appeal to prosocial sentiments might have more to gain than those that appeal to threats.

References

Arenas, A., Cota, W., Gomez-Gardenes, J., Gomez, S., Granell, C., Matamalas, J. T., . . . Steinegger, B. (2020). Derivation of the effective reproduction number R for COVID-19 in relation to mobility restrictions and confinement. *medRxiv*, 2020.2004.2006.20054320. doi:10.1101/2020.04.06.20054320

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38*(1), 23-32. doi:https://doi.org/10.1016/0022-3999(94)90005-1

Bagozzi, R. P., & Pieters, R. (1998). Goal-directed Emotions. *Cognition and Emotion, 12*(1), 1-26. doi:10.1080/026999398379754

Baron, J. (1994). *Thinking and deciding, 2nd ed*. New York, NY, US: Cambridge University Press.

Beech, H. (2020). Tracking the coronavirus: How crowded asian cities tackled an epidemic. Retrieved from https://www.nytimes.com/2020/03/17/world/asia/coronavirus-singapore-hong-kong-taiwan.html

Boster, F. J., & Mongeau, P. (1984). Fear-Arousing Persuasive Messages. *Annals of the International Communication Association, 8*(1), 330-375. doi:10.1080/23808985.1984.11678581

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet, 395*(10227), 912-920. doi:https://doi.org/10.1016/S0140-6736(20)30460-8

Canipe, C. (2020). The social distancing of America. Retrieved from https://graphics.reuters.com/HEALTH-CORONAVIRUS/USA/qmypmkmwpra/

Canli, T., Sivers, H., Whitfield, S. L., Gotlib, I. H., & Gabrieli, J. D. (2002). Amygdala response to happy faces as a function of extraversion. *Science, 296*(5576), 2191. doi:10.1126/science.1068749

Carey, R. N., & Sarma, K. M. (2016). Threat appeals in health communication: messages that elicit fear and enhance perceived efficacy positively impact on young male drivers. *BMC public health, 16*, 645-645. doi:10.1186/s12889-016-3227-2

Carleton, R. N., Norton, M. A. P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders, 21*(1), 105-117. doi:https://doi.org/10.1016/j.janxdis.2006.03.014

Carretié, L., Mercado, F., Tapia, M., & Hinojosa, J. A. (2001). Emotion, attention, and the 'negativity bias', studied through event-related potentials. *International Journal of Psychophysiology, 41*(1), 75-85. doi:https://doi.org/10.1016/S0167-8760(00)00195-1

Coibion, O., Gorodnichenko, Y., & Weber, M. (2020). Labor Markets During the COVID-19 Crisis: A Preliminary View. *National Bureau of Economic Research Working Paper Series, No. 27017*. doi:10.3386/w27017

Cooper, K. E., & Nisbet, E. C. (2016). Green Narratives: How Affective Responses to Media Messages Influence Risk Perceptions and Policy Preferences About Environmental Hazards. *Science Communication, 38*(5), 626-654. doi:10.1177/1075547016666843

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology, 38*(4), 668-678. doi:10.1037/0022-3514.38.4.668

Dillard, J. P., Li, R., Meczkowski, E., Yang, C., & Shen, L. (2016). Fear Responses to Threat Appeals: Functional Form, Methodological Considerations, and Correspondence Between Static and Dynamic Data. *Communication Research, 44*(7), 997-1018. doi:10.1177/0093650216631097

Dillard, J. P., & Nabi, R. L. (2006). The Persuasive Influence of Emotion in Cancer Prevention and Detection Messages. *Journal of Communication, 56*(s1), S123-S139. doi:10.1111/j.1460-2466.2006.00286.x

Dillard, J. P., Shen, L., & Vail, R. G. (2007). Does Perceived Message Effectiveness Cause Persuasion or vice Versa? 17 Consistent Answers. *Human Communication Research, 33*(4), 467-488. doi:10.1111/j.1468-2958.2007.00308.x

Dillard, J. P., Weber, K. M., & Vail, R. G. (2007). The Relationship Between the Perceived and Actual Effectiveness of Persuasive Messages: A Meta-Analysis With Implications for Formative Campaign Research. *Journal of Communication, 57*(4), 613-631. doi:10.1111/j.1460-2466.2007.00360.x

Duhachek, A., Agrawal, N., & Han, D. (2012). Guilt Versus Shame: Coping, Fluency, and Framing in the Effectiveness of Responsible Drinking Messages. *Journal of Marketing Research, 49*(6), 928-941.

Eysenck, H. J., & Eysenck, S. B. G. (1991). *Manual of the Eysenck Personality Scales (EPS Adult), Comprising the EPQ-revised ((EPQ-R) (including Addiction and Criminality Scales) EPQ-R Short Scale Impulsiveness (IVE) Questionnaire (impulsiveness/venturesomeness/empathy)*: Hodder and Stoughton.

Feldman, L., & Hart, P. S. (2015). Using Political Efficacy Messages to Increase Climate Activism: The Mediating Role of Emotions. *Science Communication, 38*(1), 99-127. doi:10.1177/1075547015617941

Fisher, M., & Sang-Hun, C. (2020). How south korea flattened the curve. Retrieved from https://www.nytimes.com/2020/03/23/world/asia/coronavirus-south-korea-flatten-curve.html

Fitzpatrick, J., & DeSalvo, K. (2020). Helping public health officals combat COVID-19. Retrieved from https://www.blog.google/technology/health/covid-19-community-mobility-reports

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture [Press release]

Gibbons, F. X., Gerrard, M., Ouellette, J. A., & Burzette, R. (1998). Cognitive antecedents to adolescent health risk: Discriminating between behavioral intention and behavioral willingness. *Psychology & Health, 13*(2), 319-339. doi:10.1080/08870449808406754

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J Pers Soc Psychol, 85*(2), 348-362. doi:10.1037/0022-3514.85.2.348

Haas, B. W., Constable, R. T., & Canli, T. (2008). Stop the sadness: Neuroticism is associated with sustained medial prefrontal cortex response to emotional facial expressions. *Neuroimage, 42*(1), 385-392. doi:https://doi.org/10.1016/j.neuroimage.2008.04.027

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology, 14*(2), 233-238. doi:https://doi.org/10.1016/j.conb.2004.03.010

Hart, P. S., & Feldman, L. (2014). Threat Without Efficacy? Climate Change on U.S. Network News. *Science Communication, 36*(3), 325-351. doi:10.1177/1075547013520239

Hartley, C. A., & Phelps, E. A. (2010). Changing Fear: The Neurocircuitry of Emotion Regulation. *Neuropsychopharmacology, 35*(1), 136-146. doi:10.1038/npp.2009.121

Hastings, G., Stead, M., & Webb, J. (2004). Fear appeals in social marketing: Strategic and ethical reasons for concern. *Psychology & Marketing, 21*(11), 961-986. doi:10.1002/mar.20043

Heffner, J., Son, J.-Y., & FeldmanHall, O. (under revision). Emotion prediction errors guide socially adaptive behavior.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., . . . Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science, 366*(6472), 1517. doi:10.1126/science.aaw8160

Janis, I., & Feshbach, S. (1953). Effects of fear-arousing communications. *The Journal of Abnormal and Social Psychology, 48*(1), 78-92. doi:10.1037/h0060732

John Hopkins University. (2020). COVID-19 Map. Retrieved from https://coronavirus.jhu.edu/map.html

Jordan, J., Yoeli, E., & Rand, D. (2020). Don't get it or don't spread it? Comparing self-interested versus prosocially framed COVID-19 prevention messaging. doi:10.31234/osf.io/yuq7x

Kassam, K. S., & Mendes, W. B. (2013). The effects of measuring emotion: physiological reactions to emotional situations depend on whether someone is asking. *PLoS One, 8*(7), e64959-e64959. doi:10.1371/journal.pone.0064959

Kawohl, W., & Nordt, C. (2020). COVID-19, unemployment, and suicide. *The Lancet Psychiatry, 7*(5), 389-390. doi:https://doi.org/10.1016/S2215-0366(20)30141-3

Kehoe, E. G., Toomey, J. M., Balsters, J. H., & Bokde, A. L. W. (2012). Personality modulates the effects of emotional arousal and valence on brain activation. *Social cognitive and affective neuroscience, 7*(7), 858-870. doi:10.1093/scan/nsr059

Kelly, B. J., & Hornik, R. C. (2016). Effects of Framing Health Messages in Terms of Benefits to Loved Ones or Others: An Experimental Study. *Health Communication, 31*(10), 1284-1290. doi:10.1080/10410236.2015.1062976

Kensinger, E. A., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proc Natl Acad Sci U S A, 101*(9), 3310. doi:10.1073/pnas.0306408101

Krisher, H. P., Darley, S. A., & Darley, J. M. (1973). Fear-provoking recommendations, intentions to take preventive actions, and actual preventive actions. *Journal of Personality and Social Psychology, 26*(2), 301-308. doi:10.1037/h0034465

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lang, A., & Yegiyan, N. S. (2008). Understanding the interactive effects of emotional appeal and claim strength in health messages. *Journal of Broadcasting & Electronic Media, 52*(3), 432-447. doi:10.1080/08838150802205629

Leiserowitz, A. (2006). Climate Change Risk Perception and Policy Preferences: The Role of Affect, Imagery, and Values. *Climatic Change, 77*(1), 45-72. doi:10.1007/s10584-006-9059-9

Leventhal, H. (1970). Findings and Theory in the Study of Fear Communications. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 5, pp. 119-186): Academic Press.

Lewis, I. M., Watson, B., White, K. M., & Tay, R. (2007). Promoting public health messages: Should we move beyond fear-evoking appeals in road safety? *Qual Health Res, 17*(1), 61-74. doi:10.1177/1049732306296395

Li, M., Taylor, E. G., Atkins, K. E., Chapman, G. B., & Galvani, A. P. (2016). Stimulating Influenza Vaccination via Prosocial Motives. *PLoS One, 11*(7), e0159780. doi:10.1371/journal.pone.0159780

Lunn, P., Belton, C., Lavin, C., McGowan, F., Timmons, S., & Robertson, D. (2020). Using behavioural science to help fight the coronavirus. ESRI Working Paper No. 656 March 2020. In.

Mervosh, S., Lu, D., & Swales, V. (2020). See which states and cities have told residents to stay at home. Retrieved from https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html

Monahan, J. L. (1995). Thinking positively: Using positive affect when designing health messages. In E. Maibach & R. L. Parrott (Eds.), *Designing health messages: Approaches from communication theory and public health practice* (pp. 81-98). Thousand Oaks, CA: SAGE Publications, Inc.

Moser, S. C. (2010). Communicating climate change: history, challenges, process and future directions. *WIREs Climate Change, 1*(1), 31-53. doi:10.1002/wcc.11

Myers, T. A., Nisbet, M. C., Maibach, E. W., & Leiserowitz, A. A. (2012). A public health frame arouses hopeful emotions about climate change. *Climatic Change, 113*(3), 1105-1112. doi:10.1007/s10584-012-0513-6

Nabi, R. L. (1999). A Cognitive-Functional Model for the Effects of Discrete Negative Emotions on Information Processing, Attitude Change, and Recall. *Communication Theory, 9*(3), 292-320. doi:10.1111/j.1468-2885.1999.tb00172.x

Nabi, R. L. (2002). The theoretical versus the lay meaning of disgust: Implications for emotion research. *Cognition and Emotion, 16*(5), 695-703. doi:10.1080/02699930143000437

Nabi, R. L. (2007). Emotion and persuasion: A social cognitive perspective [Lawrence Erlbaum Associates Publishers]. Retrieved

Nabi, R. L., Gustafson, A., & Jensen, R. (2018). Framing Climate Change: Exploring the Role of Emotion in Generating Advocacy Behavior. *Science Communication, 40*(4), 442-468. doi:10.1177/1075547018776019

Nisbet, M. C. (2009). Communicating Climate Change: Why Frames Matter for Public Engagement. *Environment: Science and Policy for Sustainable Development, 51*(2), 12-23. doi:10.3200/ENVT.51.2.12-23

O'Neill, S., & Nicholson-Cole, S. (2009). “Fear Won't Do It”: Promoting Positive Engagement With Climate Change Through Visual and Iconic Representations. *Science Communication, 30*(3), 355-379. doi:10.1177/1075547008329201

O’Keefe, D. J. (2012). Persuasion. In D. W (Ed.), *The international encyclopedia of communication*.

Ojala, M. (2012). Hope and climate change: the importance of hope for environmental engagement among young people. *Environmental Education Research, 18*(5), 625-642. doi:10.1080/13504622.2011.637157

Perugini, M., & Bagozzi, R. P. (2001). The role of desires and anticipated emotions in goal-directed behaviours: Broadening and deepening the theory of planned behaviour. *British Journal of Social Psychology, 40*(1), 79-98. doi:10.1348/014466601164704

Petty, R. E., & Cacioppo, J. T. (1996). *Attitudes and persuasion: Classic and contemporary approaches*. Boulder, CO, US: Westview Press.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). Linear and nonlinear mixed effects models. Retrieved from https://CRAN.R-project.org/package=nlme

Prolific Team. (2019). Representative samples on prolific. Retrieved from https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-Samples-on-Prolific

Pueyo, T. (2020). Coronavirus: Why you must act now. Retrieved from https://medium.com/@tomaspueyo/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca

Qualls, N. L., Levitt, A. M., Kanade, N., Wright-Jegede, N., Dopson, S., Biggerstaff, M., . . . Uzicanin, A. (2017). Community mitigation guidelines to prevent pandemic influenza — United States, 2017. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports, 66*(1). Retrieved from https://stacks.cdc.gov/view/cdc/45220

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population [Press release]

Reisberg, D., & Heuer, F. (1992). Remembering the details of emotional events. In *Affect and accuracy in recall: Studies of "flashbulb" memories.* (pp. 162-190). New York, NY, US: Cambridge University Press.

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review, 5*(4), 296-320. doi:10.1207/S15327957PSPR0504_2

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805-819. doi:10.1037/0022-3514.76.5.805

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502. doi:10.1037/0022-3514.57.3.493

Rusting, C. L., & Larsen, R. J. (1997). Extraversion, neuroticism, and susceptibility to positive and negative affect: A test of two theoretical models. *Personality and Individual Differences, 22*(5), 607-612. doi:https://doi.org/10.1016/S0191-8869(96)00246-2

Sauer, C., Auspurg, K., & Hinz, T. (2020). Designing Multi-Factorial Survey Experiments: Effects of Presentation Style (Text or Table), Answering Scales, and Vignette Order. *methods, data, analyses, 14*(2), 20.

Shen, L. (2011). The Effectiveness of Empathy- Versus Fear-Arousing Antismoking PSAs. *Health Communication, 26*(5), 404-415. doi:10.1080/10410236.2011.552480

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69-81. doi:https://doi.org/10.1016/j.jrp.2017.02.004

Spitzer, R. L., Kroenke, K., Williams, J. B., & Lowe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med, 166*(10), 1092-1097. doi:10.1001/archinte.166.10.1092

Storbeck, J., & Clore, G. L. (2008). Affective Arousal as Information: How Affective Arousal Influences Judgments, Learning, and Memory. *Social and Personality Psychology Compass, 2*(5), 1824-1843. doi:10.1111/j.1751-9004.2008.00138.x

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological bulletin, 141*(6), 1178-1204. doi:10.1037/a0039729

Tay, R. S., & Ozanne, L. (2002). Who are we scaring with high fear road safety advertising campaigns. *Asia Pacific Journal of Transport, 4*(Summer), 1-12. Retrieved from https://eprints.qut.edu.au/7018/

The Associated Press. (2020). Trump resists national shutdown, leaving it up to states. Retrieved from https://www.nytimes.com/aponline/2020/04/01/us/politics/ap-us-virus-outbreak-national-shutdown.html

The White House. (2020). Proclamation on declaring a national emergency concering the novel coronavirus disease (COVID-19) outbreak. Retrieved from https://www.whitehouse.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/

The White House, & Centers for Disease Control and Prevention. (2020). The president's coronavirus guidelines for America. Retrieved from https://www.whitehouse.gov/briefings-statements/coronavirus-guidelines-america/

Williams, W. C., Morelli, S. A., Ong, D. C., & Zaki, J. (2018). Interpersonal emotion regulation: Implications for affiliation, perceived support, relationships, and well-being [American Psychological Association doi:10.1037/pspi0000132]. Retrieved

Witte, K., & Allen, M. (2000). A Meta-Analysis of Fear Appeals: Implications for Effective Public Health Campaigns. *Health Education & Behavior*, *27*(5), 591-615. doi:10.1177/109019810002700506

World Health Organization. (2020). Coronavirus disease (COVID-19) advice for the public. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

Zeelenberg, M., & Pieters, R. (2006). Feeling Is for Doing: A Pragmatic Approach to the Study of Emotions in Economic Behavior. In *Social psychology and economics.* (pp. 117-137). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Chapter 4

Emotion prediction errors guide socially adaptive behaviour

Joseph Heffner, Jae-Young Son, & Oriel FeldmanHall
Published in Nature Human Behavior (2021)

Introduction

How do we learn to make adaptive decisions, such as whether to avoid a risky financial endeavor or start a collaboration with a new colleague? A rich literature on value-based decision-making illustrates that choices are made based on the expectation of rewards, and that violations of these reward expectations—that is, prediction errors (PEs)—enable an agent to update their knowledge about their environment to facilitate survival (B. King-Casas et al., 2005; Pessiglione et al., 2006; W. Schultz et al., 1997; Schultz & Dickinson, 2000). Over the last few decades, these insights have been elegantly encapsulated in a reinforcement learning framework (Sutton & Barto, 1998), which has served as the foundation for virtually all standard models of decision-making. Even complex social behaviors, such as affiliating with coworkers or reconciling with a spouse, are thought to be motivated by the violation of expected outcomes (Ho et al., 2017; Ruff & Fehr, 2014). To illustrate, a colleague's failure to meet a deadline might generate a negative PE, which in turn drives learning through continued reinforcement (for example, this colleague is often late to meetings) and adjustments of future behaviors (for example, collaborations with this colleague are to be avoided).

In parallel with research linking reward to decision-making, a separate literature also demonstrates that emotion exerts a powerful influence on choice (Bechara et al., 1997; Dalgleish, 2004; Phelps, 2006; Phelps et al., 2014; Vuilleumier, 2005). Although there has been interest in understanding how anticipated emotions affect behavior (Caplin & Leahy, 2001; Knutson & Greer, 2008; Mellers et al., 1999; Mellers & McGraw, 2001; Nielsen et al., 2008; Richard et al., 1996), relatively little work has examined how the violation of expected emotions—a concept we label emotion PEs—influence decision-making, especially in the context of social interactions (FeldmanHall & Chang, 2018; Hétu et al., 2017b; Xiang et al., 2013). A person may, for example, avoid collaborating because she dreads interactions with her aloof colleague, only to find out that once the collaboration begins, her colleague is actually quite warm and humorous. The unexpected joy of working with this colleague therefore produces a positive emotion PE, which motivates more extensive future collaborations. Prior work shows that sophisticated mental models of emotion are used to predict how other people transition between distinct emotional states (Thornton & Tamir, 2017), and predictions about expected aversive emotions such as regret and guilt can shape social interactions (Chang & Sanfey, 2009; Chang & Sanfey, 2011; Loewenstein & Lerner, 2003). However, whether violations of expected emotions also affect decision-making is an open question (Chang & Eshin, 2018; FeldmanHall & Chang, 2018).

Additionally, little is known about how reward and emotion relate to one another, as past work on decision-making has either ignored emotional experiences or assumed that emotion is synonymous with reward value (Dolan, 2002; Hartley & Sokol-Hessner, 2017; O'Doherty et al., 2001; Pessoa, 2009; Sanfey, 2007; Schultz & Dickinson, 2000). For example, in a reinforcement learning framework, external rewards (for example, money, juice) are used to update an agent's value function, and any state changes (such as emotions) are considered to be nuisance variables (Juechems & Summerfield, 2019). Other accounts hint that emotions are simply an internal proxy for value, such that emotions may shape how an individual processes the subjective value of a choice by applying a (nonlinear) transformation to objective reward (Kahneman & Tversky, 1979). This lack of consensus and clarity impacts the specificity of theories of decision-making and hampers insight into a variety of psychopathologies that are canonically associated with deficits in both reward and emotion processing (Hooker & Park, 2002; Ironside et al., 2019; Treadway & Zald, 2013; Whitton et al., 2015). For instance, it has not been determined whether emotion and reward independently or jointly impact socially maladaptive behaviors accompanying mood disorders, such as depression (Pessoa, 2008). Therefore, to gain a holistic understanding of the mechanisms guiding adaptive social decision-making (Berridge, 2007; Berridge et al., 2009), it is critical to map the relative contributions of reward and emotion PEs to behaviour.

To test how strongly reward and emotion PEs impact social behaviours such as punishing or forgiving others, we quantify how violations of emotional expectations bias choices in multiple interactive economic games. As a direct analogue to reward PEs, we examine emotion PEs using a framework that treats emotion by its basic psychological constituents free of any implied cognitive structures (Russell, 2003). This model of emotion partitions emotional experiences into the affective dimensions of valence (pleasurableness) and arousal (alertness/action; Russell & Barrett, 1999; Russell et al., 1989), which jointly constitute the core affect of emotion (we have adopted the term ‘emotion PEs’, rather than affective PEs, as it captures the conscious emotional experiences participants are being asked to measure and report during these tasks.). We developed a technique that measures real-time fluctuations in emotions as the decision-process unfolds, enabling us to precisely and mathematically map the subjective experience of emotion alongside economic rewards during social exchanges. This allows us to test the possibility that violations of emotion expectations influence socially consequential choices, such as deciding to punish or forgive a norm violator. In the tradition of reinforcement learning, we consider reward as an external reinforcer, such as money or food, that encourages similar future behaviors (Berridge, 2012; Daw et al., 2005; Dayan & Daw, 2008; Juechems & Summerfield, 2019; Kaelbling et al., 1996; Montague et al., 1996; Wolfram Schultz et al., 1997). This allows us to compare the relative strengths of reward and emotion PEs on choice, while remaining agnostic about whether (and/or how) these PEs may eventually be integrated into a common value signal reflecting 'net value' (Levy & Glimcher, 2012).

Across four separate experiments, participants (N=1,016) played one of two behavioral economic games—the Ultimatum Game (UG; Güth et al., 1982) or Justice Game (JG; FeldmanHall et al., 2014)—while simultaneously rating their affective experiences using a measure we term 'dynamic affective representation mapping' (dARM), adapted from the affect grid used in past research (Russell et al., 1989). This measure represents a subjective map of emotional responses where the horizontal axis characterizes the valence dimension, and the vertical axis characterizes the arousal dimension. A person who is feeling angry might, for example, report high arousal and negative valence by rating their emotional state in the upper-left corner of the grid (Fig. 1a).

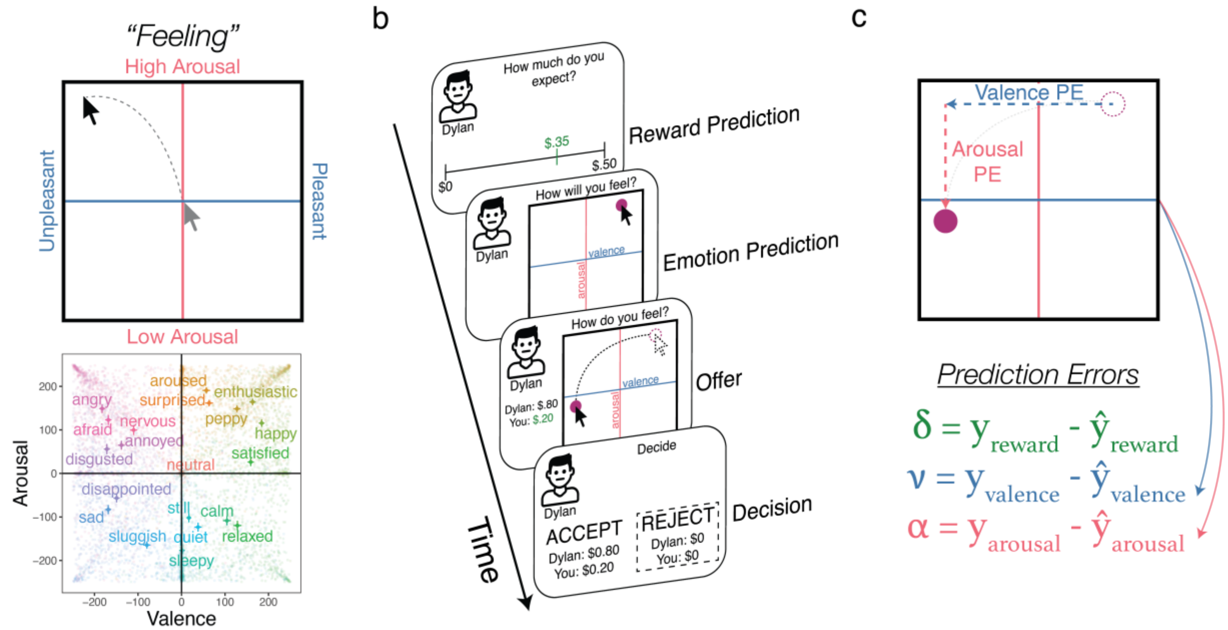


Figure 1. The tasks and PE calculations. a, *The emotion classification task. Participants rate a series of 20 feelings on the dARM measure, a 500 x 500 pixel grid only delineated by a horizontal (valence axis) and vertical line (arousal axis) along with their labels. The graph below the grid shows the average ratings for 20 feeling words all participants rated in experiment 1 (each semi-transparent data point reflects one individual rating). Error bars reflect 95% confidence intervals.* ***b,*** *The UG trial design. dARM is used in conjunction with the UG to capture emotion expectations and experiences. On each trial, participants make a prediction about how much money they expect to be offered, as well as a prediction about the emotions they expect to experience. Upon seeing the actual offer, participants report their current emotional experience. Finally, participants decide to either accept or reject the offer.* ***c,*** *Calculating reward and emotion PEs. We calculate three trial-level empirical PEs: a reward PE (δ), a valence PE (ν), and an arousal PE (α). In the equations, $\hat{y}$ refers to an individual's prediction about the reward or emotion they would experience, and y refers to their actual experience.*

In Experiment 1, participants (N=364) completed multiple rounds of a one-shot UG online, which captures punitive responses to fairness violations in a dyadic social interaction. Using a between-subjects design, participants played either as the responder or a third-party making decisions on behalf of an anonymous responder. In the UG, the responder received an unfair monetary split from the proposer, and participants were then tasked with deciding whether to accept the proposer's offer as-is or to reject the offer (that is, costly punishment such that neither the proposer nor responder receives any money). In our modified version, participants made ratings on the dARM at two critical time points: first, at the beginning of the trial before there was any monetary offer from the proposer, which captures participants' emotion expectations, and second, after the proposer makes an offer, which captures emotion experience (Fig. 1b). By using the dARM to measure emotions as a social interaction unfolds, we can mathematically compute the difference between emotion expectations and compare them to the actual emotional experience, effectively capturing emotion PEs (Methods). These emotion PEs were measured on two dimensions, valence and arousal, such that a valence PE would be calculated by the difference between the predicted versus experienced (un)pleasantness of the offer, while an arousal PE would be the difference between predicted and experienced arousal (Fig. 1c).

Mirroring how reward prediction errors are typically treated in the literature (Rutledge et al., 2014; Xiang et al., 2013), the effects of reward PEs were captured by having participants make trial-by-trial predictions about the reward they expected to receive from the Proposer, which could then be compared to the actual offer received (by subtracting the received offer from the predicted offer). This design critically allowed us to distinguish between the contributions of reward and emotion PEs during a dynamic social interaction using the following conceptual model:

$$Choice \sim Reward\ PE\ + Valence\ PE + Arousal\ PE \tag{1}$$

Before examining how well PEs for reward, valence, and arousal govern decisions to punish, we scaled all PEs at the group level prior to being modelled (similar to *z* scoring without mean-centering, as zero is the meaningful case where predictions perfectly match experience), which permitted a direct comparison of their relative contributions to choice using a common metric (Aiken & West, 1991; Iacobucci et al., 2016).

Results

Emotion and reward prediction errors have distinguishable contributions to choice

Results reveal that all three types of PEs contribute to decisions to punish. Participants punished at higher rates when experiencing less reward or valence than expected, or more arousal than expected (Table 1, Fig. 2a; the same pattern of results was found using nonparametric regression, see Supplementary Table 3 and Supplementary Figure 3). A likelihood ratio test demonstrated that the sequential addition of valence ($\chi^2(4) = 512.65, p < 0.001$) and arousal PEs ($\chi^2(4) = 70.24, p < .001$) significantly improved the explanatory power of the model over a more traditional analysis which only included reward PEs. Given past work that has characterized emotional valence as a byproduct of reward processing (Delgado et al., 2003; Lang & Bradley, 2008; Murray, 2007), it is particularly noteworthy that we find a unique contribution of valence PEs—for example,, surprisingly negative feelings such as disappointment or sadness—for punitive decisions. While the valence and reward PEs are correlated at the intra-individual level ($r_{rm} = 0.80$, $p < .001$), the variance inflation factor (VIF) statistics indicate low collinearity between these predictors in our model ($VIF_{valence} = 1.55$, $VIF_{reward} = 1.54$, $VIF_{arousal} = 1.04$), and therefore produce reliable estimates of how strongly these PEs affect choice. Moreover, using a beta coefficient test (Clogg et al., 1995), we found that valence PEs have a significantly stronger impact on motivating punitive choices than reward PEs ($z = -3.74$, $p < .001$). That is, while people do rely on reward PEs to inform their choices, they rely even more on negative deviations from expected emotional valence.

An alternative explanation of these findings is that emotion PEs are merely soaking up the additional variance that would be typically captured by modeling individual differences in the subjective valuation of reward. To test for this possibility, we pitted our empirical PE model against standard utility models that leverage an exponential scaling parameter to capture any non-linear valuations of rewards (Kahneman & Tversky, 1979; Rutledge et al., 2014; Sokol-Hessner et al., 2009). We used the following equations to transform objective reward magnitudes into subjective value before calculating the resulting subjective reward prediction error (sRPE):

$$sRPE = reward_{actual}^{\lambda} - reward_{prediction}^{\lambda} \tag{2}$$

$$where\ 0 \leq \lambda \leq 1$$

By bounding λ, this utility model captures the diminishing marginal utility of reward in the tradition of classic utility models (Kahneman & Tversky, 1979). We incorporated sRPEs into a utility model by adding an additional free parameter w_1, which specifies the subject's weight on model-derived sRPEs:

$$\begin{gathered} utility_{accept} = w_1 sRPE \\ utility_{reject} = 0 \end{gathered} \tag{2.1}$$

The choice rule was computed by placing the utility values for each decision into a softmax function:

$$p(accept) = \frac{e^{\beta(utility_{accept})}}{e^{\beta(utility_{accept})} + e^{\beta(utility_{reject})}} \tag{2.2}$$

$$p(reject) = 1 - p(accept)$$

Thus, we can compare how our empirically-derived prediction errors (reward, valence, arousal) fare against subjective reward prediction errors that incorporate non-linear valuation of reward

(Supplementary Fig. 4 and Supplementary Table 8). Results reveal that the PE model that includes empirical reward and valence (but not arousal) PEs outperforms all other models, including those that rely on model-derived reward PEs, $t(363) = -15.27$, $p < .001$, or other impoverished models that do not account for valence PEs. Furthermore, these results suggest that valence PEs in particular are not merely reflecting individual differences in the subjective valuation of reward.

While the additive model (equation 1) provides the most direct comparison of the strength of each PE on decisions to punish, we conducted a secondary analysis to assess whether PEs also exert any joint influence on choice. By testing all possible interactions between all three PE types in a mixed-effects regression, we found a significant three-way interaction between reward, valence, and arousal PEs ($\beta = -0.37$, SE = 0.07, $z = -4.91$, $p < .001$), a significant interaction between valence and arousal PEs ($\beta = -0.36$, SE = 0.11, $z = -3.22$, $p = .001$) and reward and arousal PEs ($\beta = -0.30$, SE = 0.11, $z = -2.65$, $p = .008$), but not between reward and valence PEs ($p = .60$; see Supplementary Table 4). Together, this suggests that the strength of a given PE is partially modulated by other PEs, such that arousal PEs appear to augment the role of valence and reward PEs.

We conducted follow-up analyses to assess the robustness of our results and check for potential non-linearities in the data. First, we tested for nonlinear effects of PEs on decision-making using a generalized additive mixed effects model, which showed that the marginal contribution of valence PEs had a stronger unique contribution to choice than reward PEs (see Supplementary Fig. 3 and Supplementary Table 3). Second, we tested the strength of prediction errors by controlling for expectations (that is, modelling both the prediction and PE in the same regression;

Rutledge et al., 2014), which revealed that valence and reward PEs still explain significant variance in decisions to punish, even when controlling for expectations (see Supplementary Table 5). Third, we can directly examine the contributions of rewarding and emotional experiences on decisions to punish by fitting a model that only includes information about participants' actual experiences of reward and emotion, independent of their expectations. Results reveal that experienced reward (that is, the offer itself) predicts decisions to punish more strongly than experiences of emotional valence (beta-comparison $z = -4.37$, $p < .001$) or arousal (beta-comparison $z = -10.57$, $p < .001$). Finally, since querying emotion predictions directly after reward predictions could have diminished the role of reward PEs, we ran a subsequent pre-registered experiment to replicate our findings while controlling for potential ordering effects (https://osf.io/3mgxz/). In experiment 2 (N = 228), there was no evidence that reward PEs were dampened by the presence of asking participants to predict and report their emotional experiences (see Supplementary Fig. 7 and Supplementary Tables 9 – 11). Moreover, as observed in experiment 1, valence PEs had the strongest impact on shaping decisions to punish compared to reward PEs (beta comparison: $z = -2.57$, $p = .005$).

These results have three important implications. First, reward and emotion expectation violations are distinguishable inputs during decisions to punish. Second, although reward PEs have traditionally been treated as the predominant driver of punitive decisions in social exchanges (Brooks King-Casas et al., 2005; Montague & Lohrenz, 2007), our findings instead indicate that valence emotion PEs are actually the strongest motivator. Third, when considering the direct contributions of experienced reward and emotion, reward appears to more strongly bias behavior than emotion, illustrating that emotion only outperforms reward once PEs are considered.

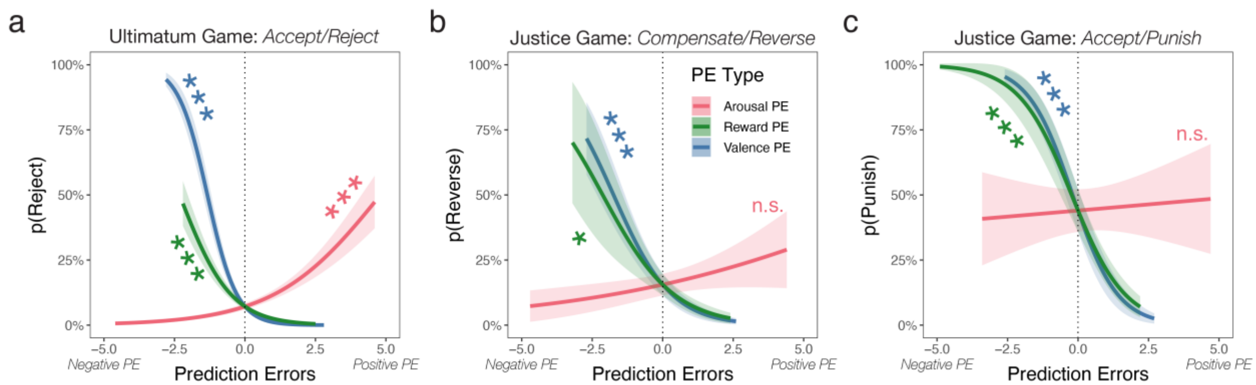


Figure 2. Emotion PEs underpin punitive behavior in the UG and JG*. *a-c, *Participants completed either the UG or JG. The lines on each graph reflect the probability of different choice pairs including rejecting vs accepting in the UG (**a**), reversing vs compensating in the JG (**b**) and punishing vs accepting in the JG (**c**). The colour of each line indicates reward (green), arousal (red), and valence (blue) PEs. Negative values reflect negative prediction errors, indicating less money (reward), less pleasantness (valence), and less arousal than expected. Shaded areas reflect ±1 s.e. ***P<0.001, **P<0.01, *P<0.05.*

Table 1. Experiment 1: valence PEs predict decisions to punish better than reward PEs

$$Punish_{i,t} \sim \beta_0 + \beta_1 Reward\ PE_{i,t} + \beta_2 Valence\ PE_{i,t} + \beta_3 Arousal\ PE_{i,t} + \varepsilon$$

Variable	Estimate (SE)	z	p
Punish			
Intercept	-2.56 (0.18)	-14.26	<.001***
Reward PE	-1.10 (0.14)	-7.62	<.001***
Valence PE	-1.92 (0.16)	-11.75	<.001***
Arousal PE	0.53 (0.09)	5.65	<.001***

Note. Reward PEs are calculated by taking the difference between the experienced and predicted reward. Valence PEs and arousal PEs are calculated by taking the difference between the experienced and predicted emotion. All variables were scaled but not mean-centered, as the zero point on each scale refers to the meaningful instance where participants' expectations matched their experience. The model includes subject-specific random intercepts and slopes for reward PE, valence PE, and arousal PE. The dataset includes 7,280 observations from 364 participants. *** p <.001.

These findings demonstrate a link between emotion prediction errors and punishment, suggesting that violations of emotion expectations are integral to motivating social choice. It remains unclear, however, how emotions are constructed during the decision-making process, or when these emotional experiences ultimately bias choice. Even when the contributions of emotion are considered alongside reward, emotions are typically treated as a static input rather than a dynamic process (Sanfey et al., 2003; Xiang et al., 2013). This assumption places artificial constraints on the role emotions play in biasing choice, and is incongruent with theoretical accounts claiming that dynamic fluctuation is a core feature of emotion (Nielsen et al., 2008). Indeed, most major theories of emotion propose that changes in the intensity of experienced affect over time can be integral in shaping behaviors (Diener et al., 1985; Scherer, 2009), and can adaptively vary depending on cue relevant environmental changes (Kuppens et al., 2010).

Temporal dynamics of emotion and its relationship with choice

Therefore, to gain a more granular perspective of how emotion biases choice, and to test the robustness of the emotion PE effect observed in experiments 1 and 2, we conducted a third experiment exploring the temporal dynamics of emotion while simultaneously employing a stronger experimental control for the influence of reward on social choice. One of the limitations of using the UG to study dynamic changes in reward and emotion is that each of the options (that is, accept or reject) results in different monetary reward. However, if the relative contribution of reward were experimentally held constant, this would allow us to decouple monetary reward from emotion, and more directly examine how emotion PEs influence choice. Therefore, to control for the variable influence of reward, in experiment 3, participants (N=73) played a modified version of the UG called the JG in a laboratory setting (FeldmanHall et al., 2014). In the JG, participants always played as the responder and received unfair offers from various proposers. After receiving an offer, responders could redistribute the money between themselves and the proposer. Analogously to the UG, two of the redistribution options were to accept (take the offer as-is) or punish (reduce proposer's payoff to match the responder's). The JG introduces two additional unique options which capture preferences between non-punitive responses and cost-free punishment: Responders could non-punitively compensate by increasing their own payout to match the amount of money proposers kept for themselves or apply a cost-free punishment by reversing the proposed payoffs, such that proposers get what they offered to responders (Methods). On each trial, only two of the available four options were randomly presented, and participants were never aware which two options would be available. This structure was important for two reasons. First, because there was uncertainty regarding which choice pair participants would receive, the final monetary outcomes were experimentally decoupled from the proposer's original offer. This provided an ideal testbed for studying continuous fluctuations in emotions, as participants' emotions could change over the course of a trial as they received new information about proposers' offers and the possible ways to redistribute the money. Second, by matching the participant's payout by holding reward constant when certain choices were presented together (that is, compensate/reverse and accept/punish), the task structure provides a strong experimental test of how emotion influences social choice independently from the final reward outcome.

As before, we used the dARM to measure participants' emotion predictions about the proposer's offer and emotional experiences after receiving the proposer's offer. Because the available choice pair was unpredictable, we additionally measured participants' emotional experiences after making their decision. All emotion measurements were sampled every 10ms using mouse tracking, which allowed us to continually measure emotion predictions and experiences as they unfolded in real time.

We first calculated emotion PEs using participants' final emotion rating (mirroring the analysis performed in experiments 1 and 2). Results reveal that negative valence emotion PEs robustly predict choices to Punish ($\beta = -1.26$, SE = .24, z = -5.18, $p < .001$) and Reverse ($\beta = -0.97$, SE =

.26, z = -3.74, $p < .001$). Although significantly predictive of punitive behavior, these valence emotion PEs did not outperform reward PEs for either choice pair (accept/punish: $z = -.52$, $p = .30$, Fig. 2B; Compensate/Reverse: $z = -.43$, $p = .34$, Fig. 2C). In contrast to Experiments 1-2, arousal PEs were not significantly predictive of decisions to Punish or Reverse (all Ps>.27, see Supplementary Tables 12 and 13), suggesting that the strength of the arousal PE signal may be context dependent. In addition, we tested how these empirically derived PEs fare against subjective (model derived, detailed above in [equation 2]) reward prediction errors. We found that a model that includes empirical reward and valence—but not arousal—PEs outperforms all other models, including those that only rely on model-derived reward prediction errors ($t(72) = -49.7$, $p < .001$; see Supplementary Figure 8 and Supplementary Table 15). Taken alongside the results from Experiments 1-2, these findings suggest that emotion prediction errors—and in particular valence PEs—play a unique role in biasing various types of social choices, are at least equally potent as reward prediction errors in motivating social behaviors, and can be more powerful than reward prediction errors in certain contexts.

To further examine how the construction of emotion biases choice, we probed real-time fluctuations in participants' emotional experiences. Participants were permitted to report their emotional experiences at their own pace, resulting in trials of different lengths. We therefore resampled all emotion trajectories to a normalized timescale consisting of 100 timepoints to aid interpretation (Freeman & Ambady, 2010), and then averaged participants' emotional responses across choices. Figure 3 shows the unique emotion trajectory across time, separately for valence and arousal, for any given decision (Accept, Punish, Reverse, Compensate). When comparing the emotional trajectories in Compensate versus Reverse trials only, results reveal that participants' eventual decisions to Reverse an unfair offer—a retributive eye-for-an-eye response—can be predicted by the 37^{th} timepoint on the normalized scale, which corresponds to 1.65s on an absolute timescale. Early emotional trajectories were so potent that we could even predict some decisions as early as half a second on an absolute timescale (for the Accept/Compensate pair with a moderately unfair offer; see Supplementary Table 16). We also examined how individuals' choices alter their emotions after the social interaction. While on average everyone's emotional valence increased after making a choice, those who responded punitively (Reverse or Punish) rapidly reported feeling positive emotional valence (Fig. 3)—a "joy of punishment" effect. Together, these results reveal that, by experimentally stripping away the potential influence of monetary reward on choice, there is a striking impact of early emotional experiences on guiding which subsequent choice is taken—which provides further evidence that reward and emotion have unique inputs to social choice.

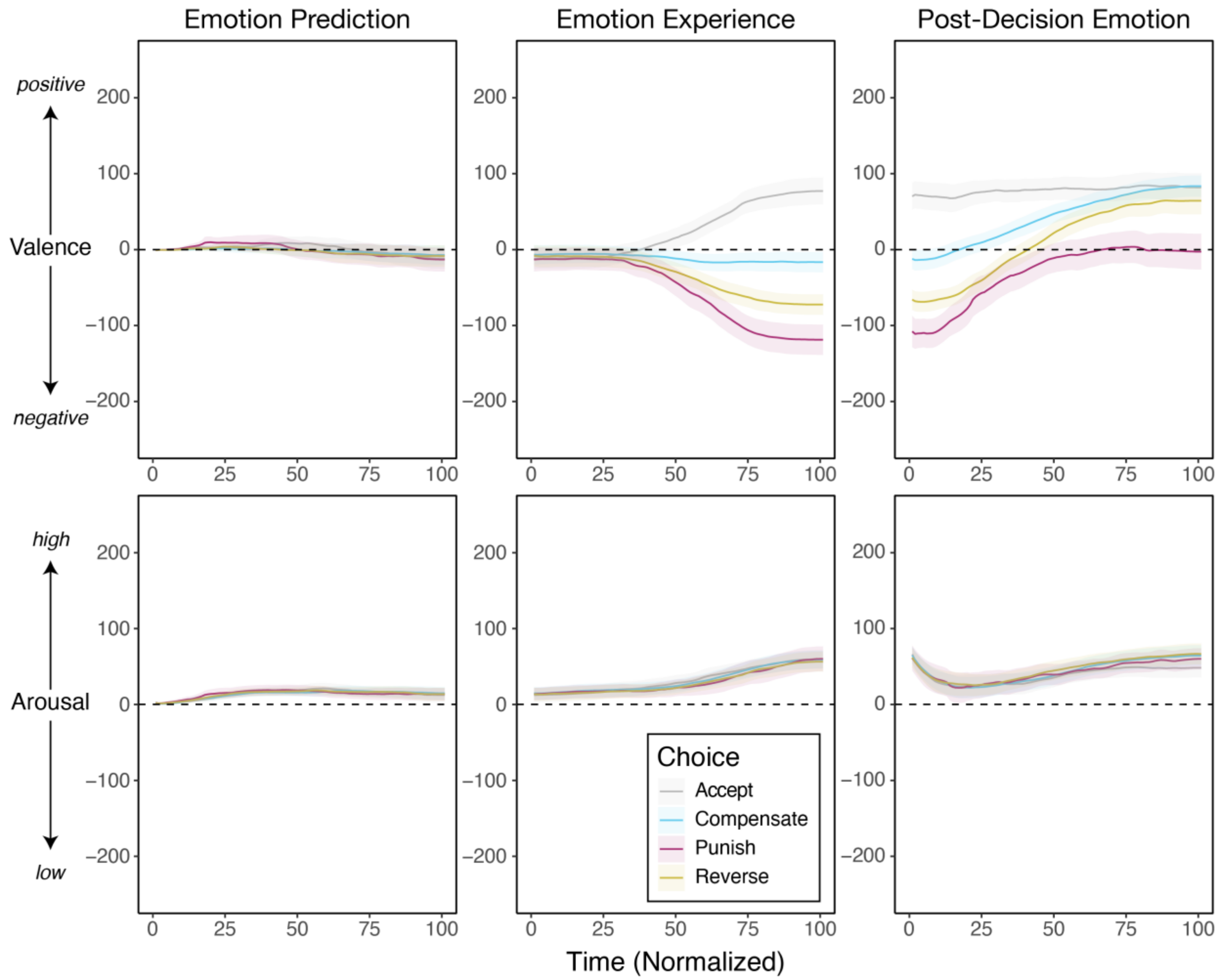


Figure 3. Temporal dynamics of emotional experiences during choice. *Participants used the dARM to continuously report their emotional experiences at every stage of the JG. For all measurements, participants' data were normalized for the time it took to make a rating, such that 1 represents the start of the trial and 101 represents the final emotion rating. The average valence and arousal are plotted over time. All shaded areas represent 95% CIs.*

Functional dissociation between reward and emotion prediction errors

The privileged role of emotion prediction errors in guiding social behavior has important implications for mood disorders such as depression, which is often characterized by impairments in both reward and emotion processing (Whitton et al., 2015). To date, however, extant research on depression has examined reward and emotion in a siloed manner (Admon & Pizzagalli, 2015; Keren et al., 2018)—that is, they have not been interrogated side-by-side within the same paradigm—and there have been few attempts to link them to decision-making in lockstep. Consequently, it remains unclear whether reward or emotion deficits are the primary contributor to symptoms and socially maladaptive behaviors seen in clinical populations, such as anhedonia (Treadway & Zald, 2013) and avolition (Dowd et al., 2016). We therefore conducted a preregistered fourth experiment (https://osf.io/qfejk/) comparing healthy controls against individuals reporting significant levels of depressive symptoms.

Experiment 4 (N=351) measured participants' predictions and experiences about reward and emotion in the UG. After completing the task, participants completed questionnaires indexing various mood disorders, including the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977). Participants were classified as at risk for depression or a healthy control based on scoring guidelines of depression symptomatology of the CES-D (see Methods for details about the CES-D and all additional measures).

We first observed that compared to healthy controls, those at risk for depression were less sensitive to the offer unfairness, which led them to be more punitive for fair offers and less punitive for unfair offers (Table 2). Given these observed behavioral differences, our primary goal was to then examine whether participants at risk for depression demonstrated aberrant use of reward and/or emotion PEs when deciding to punish. Replicating our findings from the first three experiments, healthy controls (N=205) relied most heavily on valence prediction errors when making punitive decisions ($\beta = -2.08$, SE = 0.29, z = -7.15, $p < .001$). Valence PEs were also more predictive of decisions to punish than reward PEs ($\beta = -1.41$, SE = 0.23, z = -6.03, $p =$ $< .001$; beta coefficient test $z = -1.80$, $p = .036$; Fig. 4A). In contrast, individuals at risk for depression (N=146) demonstrated no reliance on arousal PEs ($\beta = 0.01$, SE = 0.13, z = 0.09, $p =$

.93), and significantly more reliance on reward compared to valence PEs (Reward $\beta = -1.43$, SE = 0.18, z = -7.79, $p < .001$; Valence $\beta = -0.94$, SE = 0.16, z = -5.81, $p < .001$; beta coefficient test $z = -1.98$, $p = 0.02$; Fig. 4A)—which accords with previous work showing that people with depression exhibit intact reward prediction errors in certain contexts (Moutoussis et al., 2018).

Using the healthy controls as a benchmark, those at risk for depression exhibited selective impairment in their use of both emotion prediction errors when punishing, but there were no observable differences in reward prediction error processing. Remarkably, the attenuated reliance on valence and arousal prediction errors led to less punishment of a transgressor compared to healthy controls (Table 3; Fig. 4B). While it is possible that participants at risk for depression could simply have less reliable emotion prediction errors, this explanation is unlikely given the nearly identical distribution of arousal and valence prediction errors between groups (see Supplementary Figure 13). These findings reveal a functional dissociation between emotion and reward during the decision-making process in depression, suggesting that the two may be cognitively separable.

To further probe why participants at risk of depression relied less on emotion prediction errors, we next examined participants' responses in an independent emotion classification task. In this task, participants rated 20 canonical emotion labels (for example, anger) on the dARM, which required them to draw upon their past memories and knowledge of how they experience each of these emotions (Methods). Results reveal that individuals classified as depressed have a smaller range of emotional experiences (Welch two sample *t*-test, $t(263.34) = 4.33$, $p < .001$, Hedge's $g = 0.49$; Fig. 4C, 4D). Restricted emotion representations were observed along both the valence and arousal dimensions. This suggests that depression may be linked with impairments in how emotional experiences are represented (Demiralp et al., 2012; Lindquist & Barrett, 2008), and may help explain why depression attenuated the influence of emotion prediction errors on decisions to punish in our experiment.

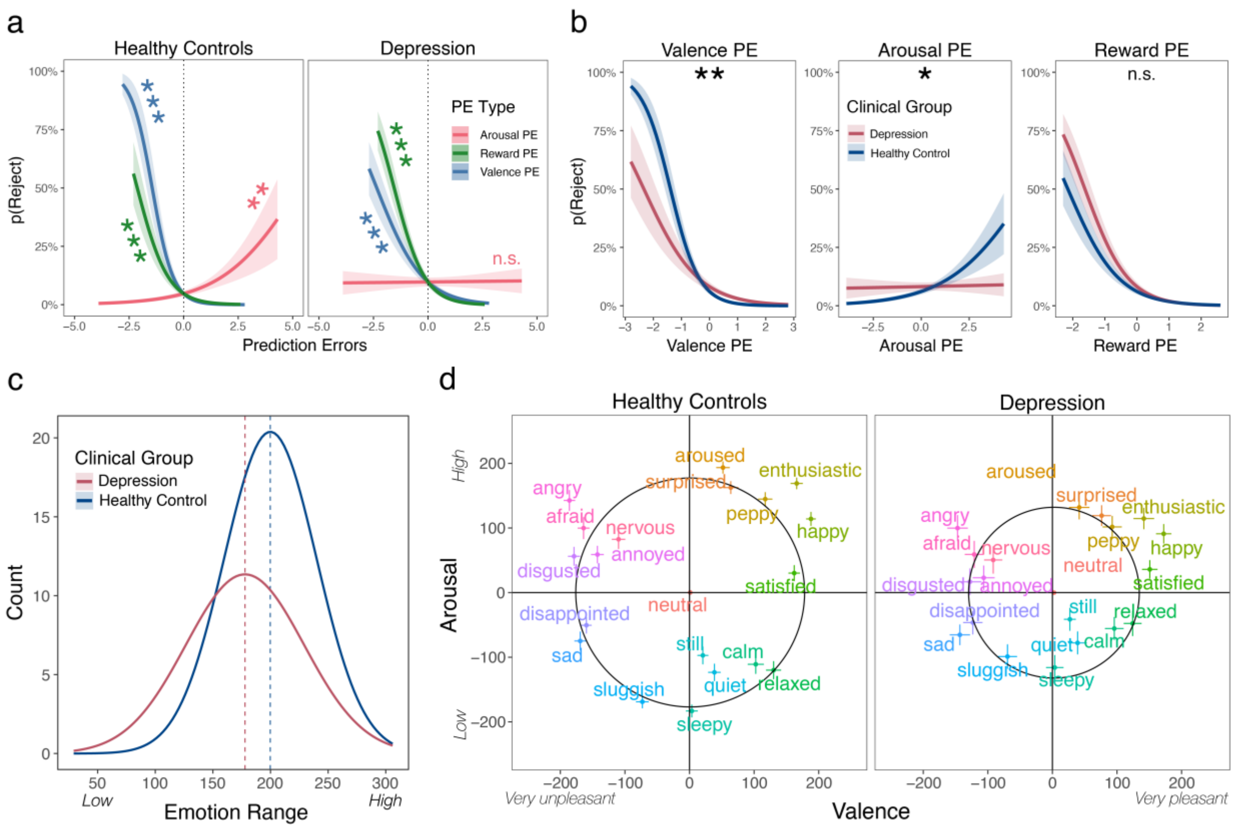


Figure 4. Results of experiment 4. a, *PE use in healthy controls and those at risk of depression. The probability of rejecting the offer is plotted for all three types of PEs (reward, arousal, and valence) for healthy controls (left) and those reporting significant levels of depression (right). Negative values reflect negative PEs, while positive values reflect positive PEs. Analyses represent separate regression effects. Shaded areas reflect ±1 SE.* ***b,*** *PEs plotted by group. The use of each PE is plotted for both healthy controls and individuals at risk of depression. Analyses represent interactions between each PE and group.* ***c,*** *Emotion range. Each participant's emotion ratings were used to calculate the average distance of their ratings from neutral (that is, the radius of their unique circumplex), thereby indexing their emotional range. Histograms represent normal distributions for both groups and dashed lines indicate their respective means.* ***d,*** *Group-level average emotion ratings. Participants rated 20 typical emotions using the dARM and the circles represent emotion range fits for each group. Error bars represent 95% CIs. Shaded areas reflect ±1 SE. p < .001***, p < .01**, p < .05*.*

Table 2. Experiment 4: individuals at risk of depression are less sensitive to unfairness

$$Punish_{i,t} \sim \beta_0 + \beta_1 Depression_i + \beta_2 Unfairness_{i,t} + \beta_3 (Depression \times Unfairness)_{i,t} + \varepsilon$$

Variable	Estimate (SE)	z	p
Punish			
Intercept	-4.57 (0.48)	-9.56	<.001***
Depressed	1.42 (0.60)	2.38	.018*
Unfairness	5.57 (0.38)	14.53	<.001***
Depressed × Unfairness	-1.80 (0.42)	-4.24	<.001***

Note. Unfairness is scaled and mean-centered. Depression is a binary variable with healthy controls (0) and those at risk of depression (1). The model includes subject-specific random intercepts and slopes for Unfairness. The dataset includes 7,020 observations from 351 participants.

* p < .05. *** p <.001.

Table 3. Experiment 4: individuals at risk of depression have selective impairment in emotion (but not reward) PEs

$$Punish_{i,t} \sim \beta_0 + \beta_1 rPE_{i,t} + \beta_2 vPE_{i,t} + \beta_3 aPE_{i,t} + \beta_4 Depression_i + \beta_5 (rPE \times Depression)_{i,t} + \beta_6 (vPE \times Depression)_{i,t} + \beta_7 (aPE \times Depression)_{i,t} + \varepsilon$$

Variable	Estimate (SE)	z	p
Punish			
Intercept	-2.73 (0.22)	-12.58	<.001***
Reward PE	-1.27 (0.19)	-6.54	<.001***
Valence PE	-1.96 (0.22)	-9.03	<.001***
Arousal PE	0.49 (0.14)	3.58	<.001***
Depression	0.31 (0.30)	1.04	.296
Reward PE × Depression	-0.22 (0.25)	-0.88	.378
Valence PE × Depression	0.93 (0.29)	3.24	.001**
Arousal PE × Depression	-0.47 (0.18)	-2.54	.011*

Note. Reward PEs are calculated by taking the difference between the experienced and predicted reward. Valence PEs and arousal PEs are calculated by taking the difference between the experienced and predicted emotion. All variables were scaled but not mean-centered, as the zero point on each scale refers to the meaningful instance where expectations match experiences. Depression is a binary variable with healthy controls (0) and those at risk of depression (1). The model includes subject-specific random intercepts and slopes for reward PE, valence PE, and arousal PE. The dataset includes 7,020 observations from 351 participants.

* p < .05. ** p < .01. *** p <.001.

Discussion

Historically, there have been two major perspectives concerning the relationship between emotion, reward, and decision-making. Either emotion has been considered irrelevant or purely incidental to choice, or else emotion and reward have been treated as so intrinsically intertwined that they cannot be disentangled, thus serving similar functional roles (Dolan, 2002; Hartley & Sokol-Hessner, 2017; Pessoa, 2009; Sanfey, 2007). Here, we examine the relationship between emotion and reward, revealing that neither of these accounts is accurate. By classifying emotions into distinct affective components (valence and arousal), we interrogated the possibility that emotion and reward prediction errors make both unique and interactive contributions to socially-consequential choices, such as punishing or forgiving moral transgressors. Our findings document that people rely heavily on violations of their emotional expectations to make social decisions, and that these emotion PEs are just as powerful, if not more potent, than reward PEs in guiding social behaviors.

By mathematically computing the consciously accessible affective states using the dARM, we were able to measure PEs for different dimensions of emotion (Russell et al., 1989), explore whether specific types of expectation violations are especially consequential for social decisions and broaden the scope of research on anticipated emotion by creating a generalizable framework that does not rely on discrete emotions such as guilt or regret (Chang et al., 2011). These findings build on a rich literature documenting how decision-making is influenced by emotion (Bechara et al., 1997; Dalgleish, 2004; Phelps, 2006; Phelps et al., 2014; Sanfey, 2007). We demonstrate that negative valence PE (negative surprise) and positive arousal PE (experiencing more arousal than expected) increase the likelihood of punishing, such that valence and arousal PEs have independent and opposite effects on choice. We note, however, that the influence of arousal PEs on choice appears context-dependent, as we did not observe an effect of arousal PEs in experiment 3. In contrast to past accounts (Sanfey, 2007; Xiang et al., 2013), these results imply emotions ought to be considered in relation to the violation of an emotional expectation—not just in the emotional experience itself.

Even when the emotional experience is considered in isolation, methodological advances measuring moment-to-moment emotional changes can document the real-time evolution of how this process unfolds, clarifying emotion's role in social decision-making. For example, we find that early emotional reactions—those that come online in less than a second during the social interaction—quickly and powerfully predict what social choices people subsequently make. Furthermore, the choices people make can drastically influence their emotional states: Choosing to punish perpetrators results in a rapid positive boost ('a joy of punishment') in the wake of their decision. Together, these results accord with a growing literature on predictive processing (Hutchinson & Barrett, 2019), and suggests that early, transient emotional states during an unfair social exchange are essential in governing whether people ultimately decide to punish or forgive a perpetrator.

By adopting a PE framework to explore how social decisions are shaped by violations of emotional expectations, we were able to compare the strength of reward and emotion prediction errors in motivating decisions to punish and help others. Foundational work in decision-making has focused on rewards as external reinforcers, illustrating that reward PE are an important mechanism for enabling adaptive behavior, as they allow people to compare their expectations against their experiences to modify actions accordingly (Brooks King-Casas et al., 2005;

Wolfram Schultz et al., 1997). We observe this in our own studies. Reward PE explain significant variation in social behaviors and critically contribute to the choices people make during social interactions. In a similar vein, people rely on violations of their emotional expectations to calibrate their choices, and we observed a particularly robust role of valence PEs in predicting social choices across contexts. Arousal PEs, on the other hand, seem to be more sensitive to the social context, as they were not universally deployed across all experiments (that is, the arousal PE effect was not observed in the JG, and once predictions were accounted for in the UG, arousal PEs no longer provided significant predictive value). Our computational model further hints that regardless of context, some individuals may not rely on arousal PEs at all to inform their choices. When taken together, our results suggest that the different types of prediction errors uniquely and interactively contribute to social choice, such that neither emotion nor reward predictions alone tell the whole story. Rather, social choices appear to be the result of joint inputs from both emotion and reward prediction errors.

These findings are compatible with the theory that value is neurally encoded as a common currency, where the value of choice options under consideration are mapped to a single scale for comparison (Glimcher, 2014; Levy & Glimcher, 2012). For example, the multiple different types of prediction errors measured in our studies—reward, valence, and arousal—may feed into an integrated value signal in the prefrontal cortex (Chib et al., 2009; Smith et al., 2010). It additionally remains unknown whether a common value representation places equal weight on each kind of prediction error, or whether value is asymmetrically biased by valence PEs. Future work can help clarify how (and where) these distinct emotion and reward PEs are processed in the brain and the extent to which they are separable.

The adaptive qualities of emotion prediction errors become readily apparent when considering people at risk for depression, who were selectively impaired in using emotion—but not reward—prediction errors when making social decisions. We observed that individuals reporting significant levels of depression exhibited attenuated use of valence PEs and did not rely on arousal PEs at all, which led to less punishment of a norm transgressor. In contrast, they exhibited fully intact use of reward prediction errors, which accords with past research (Rutledge et al., 2017) although this may be contingent on the learning context (Gradin et al., 2011; Kumar et al., 2018; Kumar et al., 2008). This pattern of relying on emotion PEs rather than reward PEs, seems central to healthy and adaptive social decision-making. Depression was associated with a reduced range of emotional responses in our studies, highlighting that emotional processes are fundamentally altered in mood disorders (Demiralp et al., 2012; Ehring et al., 2010). This emotional constraint may help explain the aberrant use of emotion prediction errors in those suffering from depression. Put simply, if a person is less sensitive in distinguishing between affective states, then they may be less able to choose appropriate actions given the social dynamics of the situation.

For close to a century, psychologists have sought to understand the essential drivers of human behavior. One successful framework, reinforcement learning, has elegantly illustrated that people make consequential decisions based on violations of expected rewards. This canon of work has laid the building blocks of how we understand human learning and decision-making. By adopting a similar approach, we reveal that violations of emotional expectations—emotion PEs—play an outsized role in guiding social behaviors. Using a variety of multimodal techniques, we document consistent evidence that these emotion prediction errors exert a strong influence on social behaviors, above and beyond reward prediction errors. Although past research has often placed reward prediction errors at the heart of decision-making, our results instead suggest that people robustly use violations of their emotional expectations to make decisions that influence both themselves and others. Contrary to conventional wisdom, we find that the only time reward plays a stronger role than emotion in decision-making is when experiences are considered in isolation from expectations. Together, these results highlight the critical importance that violations of expected emotions play, suggesting that emotional processes are just as consequential—if not more so—than violations of reward for guiding social behaviors.

Methods

Participants

Across all four experiments, participants (N = 1,225) received either monetary compensation or partial course credit, and provided informed consent in a manner approved by Brown University's Institutional Review Board under protocol 1607001555. In Experiment 1, we aimed to collect a sample of 350 participants, which exceeds the sample sizes used in similar paradigms using reward prediction errors to study decision-making in the Ultimatum Game (Hétu et al., 2017a; Xiang et al., 2013). We recruited 398 individuals online using Amazon Mechanical Turk (AMT). To protect against data contamination from bots posing as real participants (Chmielewski & Kucker, 2019), we excluded 34 individuals' data using a conservative measure of noncompliance on the emotion classification task, which involved correctly rating the 'neutral' feeling that we explicitly instructed ought to be in the center of the dARM (Supplementary Methods). This resulted in a final sample of 364 participants (172 female; mean age = 33.77, SD = 9.97). In Experiment 2, we collected 244 participants and excluded 16 individuals due to noncompliance, resulting in a final sample of 228 participants based on our preregistration (127 female; mean age = 35.30, SD = 11.8; Supplement Information). In Experiment 3, we recruited 75 individuals and excluded 2 individuals due to noncompliance, resulting in a final sample of 73 participants (39 female; mean age = 20.33, SD = 3.27), comparable to similar paradigms using the Justice Game (FeldmanHall et al., 2014). In Experiment 4, we aimed to collect a sample of 150 participants with depression using AMT, as detailed in our preregistration report, and we accordingly recruited a total of 508 participants. Using the preregistered exclusion criterion (identical to the one used in Experiment 1), we excluded 157 individuals from analysis due to noncompliance, resulting in a final sample of 351 participants (149 female; mean age = 35.13, SD = 10.21) with 205 healthy controls and 146 individuals classified as at risk of depression. After exclusions, the total number of participants was 1,016.

General procedure

In all experiments, participants used the dARM measure to rate their emotion experiences in real time during an emotion classification task and a behavioral economic game. After completing these tasks, participants responded to a series of individual measures and/or clinical battery, depending on the experiment.

Dynamic Affective Representation Mapping (dARM) measure

In Experiments 1, 2, and 4, we collected data using the Qualtrics online survey platform. Adapted from the affect grid used in past research (Russell et al., 1989), participants were presented with the dynamic Affective Representation Mapping (dARM) measure with a sampling resolution of 500 x 500 pixels, and asked to make their affective rating by clicking anywhere in the grid space. This enabled us to simultaneously capture fine-grained self-reports of both the valence and arousal dimensions. To familiarize participants with the use of the dARM, all participants first completed an emotion classification task. In this task, participants were asked to make ratings of 20 canonical emotion words on the grid (e.g., angry, sad, and surprised) in a randomized order. While this affective representation is typically inferred from pairwise similarity ratings of discrete emotions (Posner et al., 2005), simply training participants to interpret this subjective map has shown strong convergent validity with other approaches for emotion ratings(Russell et al., 1989). In order to capture real-time mouse-tracking in Experiment 3, which was run in the laboratory, we used the Psychtoolbox library in Matlab to implement the dARM with a spatial resolution of 500 x 500 pixels and a temporal resolution of 10ms sampling. All participants first completed the emotion classification task.

Tasks

In Experiment 1, participants played 20 rounds of the Ultimatum Game (UG) as either the Responder or a third-party. Since Responders and third-party deciders reacted to unfairness in similar ways, we collapsed across role for this analysis (see Supplementary Table 2). On each trial, participants were asked to answer the following questions: i) Predict how much reward the Responder would get (i.e., how much the Proposer would offer), within a range of $0 to $0.50; ii) Predict what emotions they expected to feel based on that reward; iii) Report their actual emotion experience upon receiving the offer; and iv) Decide whether to accept or reject the Proposer's offer. The unfairness of the offer was drawn from a pseudo-random uniform distribution such that participants saw the full range of fair ($0.50, $0.50) to unfair ($0.95, $0.05) offers.

In Experiment 2, participants played 20 rounds of the UG as the Responder. All participants completed two blocks in a counterbalanced order, a reward-only block and a reward+emotion block. In the reward-only block, participants only made reward predictions without any emotion predictions or emotion experience ratings. The reward+emotion block design was the same as Experiment 1.

In Experiment 3, participants played the Justice Game (JG) in the laboratory. In the JG, participants always play as the Responder and are paired with a unique anonymous Proposer on every trial, who offers the Responder a split of money. On each of the 54 trials, the two options presented to the Responder are drawn randomly from the four available options, ensuring that participants do not know what their choice set will be ahead of time. The four available options are: 1) Accept: keeps the offer as-is; 2) Punish: reduces the payout of the Proposer to what was offered to the Responder; 3) Compensate: increase the Responder's payout to match the Proposer's; and 4) Reverse: swapping the payouts. For example, using the notation of ($Proposer, $Responder), if the offer was highly unfair ($9, $1), the four options would be Accept ($9, $1), Punish ($1, $1), Compensate ($9, $9), and Reverse ($1, $9). The unfairness of the offer was generated such that low, medium, and high unfair offers were equally likely and each offer was generated using a truncated normal distribution (the Proposer kept $5.10-6.30 for low offers, $6.90-8.10 for medium offers, and $8.70-9.90 for high offers). Participants were asked to do the following on a trial-by-trial basis: i) Predict how much reward they would receive, ii) Predict what emotions they expected to feel based on that reward, iii) Report their emotional experience upon receiving the Proposer's offer, and iv) Make a decision about how to redistribute the money. Participants were also asked to report their emotional experience after making a decision, as to capture changes in their affective state depending on the available redistribution options. Because there were only 54 trials with six unique choice pairs and three levels of unfairness, time course analyses examining specific choice pairs (e.g., Compensate/Reverse) only include nine trials per subject.

In Experiment 4, participants played 20 rounds of the Ultimatum Game (UG) as the Responder. Otherwise, the UG design in Experiment 4 was identical to Experiment 1.

Post-Task Questionnaires

Following these tasks, participants completed a series of individual difference questionnaires. In Experiments 1 and 4, we collected two survey measures for use as potential covariates: the Emotion Regulation Questionnaire (Gross & John, 2003) and the 20-item Toronto Alexithymia Scale (Bagby et al., 1994). In Experiment 4, participants also completed the Center for Epidemiologic Studies Depression Scale (Radloff, 1977) and five survey measures to index how richly they experience reward and emotion: the Temporal Experience of Pleasure Scale (Gard et al., 2006), the Behavioral Inhibition and Behavioral Activation Scales (Carver & White, 1994), the Snaith-Hamilton Pleasure Scale (Snaith et al., 1995), the Apathy Evaluation Scale (Marin et al., 1991), and the 20-item Toronto Alexithymia Scale (Bagby et al., 1994). The primary measure of importance was the CES-D, as it allowed us to identify which participants were at risk of depression. Our hypotheses and predictions about all measures were preregistered and can be found in our OSF pre-registered report (https://osf.io/qfejk/). Analyses of the other measures are included in the Supplement.

Analysis

Across all experiments, we used logistic mixed-effects regressions to predict participants' decisions using the lme4 package in R (Bates et al., 2015). All prediction errors were calculated by taking the difference between the participants' experience (at the time of offer) and the participants' prediction (before the offer). To ensure that the beta coefficients from logistic regressions were comparable, we scaled (but did not mean-center) all prediction errors before entering them into the regression. We chose not to mean-center the prediction errors because zero indicates meaningful cases in which the participant's prediction matched their experience, therefore producing no error.

In Experiments 1 and 2, we used valence PEs, arousal PEs, and reward PEs to predict prosocial decisions to Accept or punitive decisions to Reject. The same regression specification was carried forward in all experiments. In Experiment 3, we emulated Experiment 1's analysis by taking the endpoints of the participants' mouse trajectories (i.e., the final valence and arousal ratings) to run separate logistic mixed-effects regressions for the Accept/Punish choice set and the Compensate/Reverse choice set. The regression for Experiment 4 additionally included terms accounting for being classified as in the depressed or healthy control group (based on CES-D scores), and the interactions between depression and all PE variants. To aid interpretation, we additionally performed separate regressions for participants who had significant levels of depressive symptoms and those who did not (i.e., a binary variable based on a CES-D threshold of 16, according to scoring guidelines).

In order to analyze real-time fluctuations in participants' emotions in Experiment 3, we discretized the time data into 10ms bins (i.e., our sampling rate). Because participants' emotional ratings were self-paced and had variable response times, we normalized each participants' response time on a trial-by-trial basis to compare across participants. Accordingly, participants' response times were rescaled from 1 (start of trial) to 101 (participant clicked response; Freeman et al., 2011; Wulff et al., 2019). Participants' valence and arousal measurements were averaged within each normalized time bin, allowing us to directly compare the distribution of valence and arousal responses for each of the choice sets (Accept/Punish and Compensate/Reverse) using one-way ANOVAs at each normalized time bin. To use a principled way of defining significant clusters of time bins, we used cluster-based permutation testing, which controls for multiple comparisons by generating null distributions of clusters that can be compared against true clusters (Barr, 2019; Frömer et al., 2018; Maris, 2012; Maris & Oostenveld, 2007). This method estimates how big clusters would be if there would no differences between the groups (e.g., decisions to Compensate or Reverse). Permutation tests assume that these observation labels are exchangeable under the null hypothesis, such that if there is no difference between the groups, then the labels can be randomly shuffled without consequence. For each randomly permutated time series, the largest cluster statistic is calculated (the summation of F-statistics for the largest temporally continuous cluster), which represents the largest cluster that could appear due to chance. After repeating this process 1,000 times (which builds a null distribution of clusters), we test whether the clusters observed in our data are greater than 95% of the clusters expect by chance. This ensures we can precisely quantify how the evolution of emotions affects later decisions to be punitive or forgiving while controlling for multiple comparisons.

References

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional reward processing in depression. *Current Opinion in Psychology*, *4*, 114-118. https://doi.org/https://doi.org/10.1016/j.copsyc.2014.12.011

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications, Inc.

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23-32. https://doi.org/https://doi.org/10.1016/0022-3999(94)90005-1

Barr, D. (2019). *clusterperm*. https://github.com/dalejbarr/clusterperm

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4 [sparse matrix methods; linear mixed models; penalized least squares; Cholesky decomposition]. *2015*, *67*(1), 48. https://doi.org/10.18637/jss.v067.i01

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293-1295. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9036851

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, *191*(3), 391-431. https://doi.org/10.1007/s00213-006-0578-x

Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35*(7), 1124-1143. https://doi.org/10.1111/j.1460-9568.2012.07990.x

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*, *9*(1), 65-73. https://doi.org/10.1016/j.coph.2008.12.014

Caplin, A., & Leahy, J. (2001). Psychological Expected Utility Theory and Anticipatory Feelings. *The Quarterly Journal of Economics*, *116*(1), 55-79. https://doi.org/10.1162/003355301556347

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319-333. https://doi.org/10.1037/0022-3514.67.2.319

Chang, L., & Sanfey, A. (2009). Unforgettable ultimatums? Expectation violations promote enhanced social memory following economic bargaining [10.3389/neuro.08.036.2009]. *Frontiers in Behavioral Neuroscience*, *3*, 36. https://doi.org/https://doi.org/10.3389/neuro.08.036.2009

Chang, L. J., & Eshin, J. (2018). Emotions as computational signals of goal error. In A. S. Fox, R. C. Lapate, A. J. Shackman, & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions*. Oxford University Press.

Chang, L. J., & Sanfey, A. G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Social cognitive and affective neuroscience*, *8*(3), 277-284. https://doi.org/10.1093/scan/nsr094

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, *70*(3), 560-572. https://doi.org/10.1016/j.neuron.2011.02.056

Chib, V. S., Rangel, A., Shimojo, S., & Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *The Journal of Neuroscience*, *29*(39), 12315. https://doi.org/10.1523/JNEUROSCI.2575-09.2009

Chmielewski, M., & Kucker, S. C. (2019). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, 1948550619875149. https://doi.org/10.1177/1948550619875149

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, *100*(5), 1261-1293. www.jstor.org/stable/2782277

Dalgleish, T. (2004). The emotional brain. *Nat Rev Neurosci*, *5*(7), 583-589. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15208700

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control [10.1038/nn1560]. *Nat Neurosci*, *8*(12), 1704-1711. https://doi.org/http://www.nature.com/neuro/journal/v8/n12/suppinfo/nn1560_S1.html

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429-453. https://doi.org/10.3758/CABN.8.4.429

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(1), 27-38. https://doi.org/10.3758/CABN.3.1.27

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Barrett, L. F., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., Deldin, P. J., Gotlib, I. H., & Jonides, J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychol Sci*, *23*(11), 1410-1416. https://doi.org/10.1177/0956797612444903

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: dimensions underlying positive and negative affect. *J Pers Soc Psychol*, *48*(5), 1253-1265. https://doi.org/10.1037//0022-3514.48.5.1253

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191-1194. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12424363

Dowd, E. C., Frank, M. J., Collins, A., Gold, J. M., & Barch, D. M. (2016). Probabilistic Reinforcement Learning in Patients With Schizophrenia: Relationships to Anhedonia and Avolition. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 460-473. https://doi.org/https://doi.org/10.1016/j.bpsc.2016.05.005

Ehring, T., Tuschen-Caffier, B., Schnulle, J., Fischer, S., & Gross, J. J. (2010). Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. *Emotion (Washington, D.C.)*, *10*(4), 563-572. https://doi.org/10.1037/a0019010

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*(1), 153-166. https://doi.org/10.1037/0022-3514.69.1.153

FeldmanHall, O., & Chang, L. J. (2018). Social Learning: Emotions Aid in Optimizing Goal-Directed Social Behavior. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-Directed Decision Making* (pp. 309-330). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-812098-9.00014-0

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nat Commun*, *5*, 5306.

Freeman, J., Dale, R., & Farmer, T. (2011). Hand in Motion Reveals Mind in Motion [Mini Review]. *Frontiers in Psychology*, *2*(59). https://doi.org/10.3389/fpsyg.2011.00059

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226-241. https://doi.org/10.3758/BRM.42.1.226

Frömer, R., Maier, M., & Abdel Rahman, R. (2018). Group-Level EEG-Processing Pipeline for Flexible Single Trial-Based Analyses Including Linear Mixed Models [10.3389/fnins.2018.00048]. *Frontiers in Neuroscience*, *12*, 48. https://www.frontiersin.org/article/10.3389/fnins.2018.00048

Gard, D. E., Gard, M. G., Kring, A. M., & John, O. P. (2006). Anticipatory and consummatory components of the experience of pleasure: A scale development study. *Journal of Research in Personality*, *40*(6), 1086-1102. https://doi.org/10.1016/j.jrp.2005.11.001

Glimcher, P. W. (2014). Chapter 20 - Value-Based Decision Making. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics (Second Edition)* (pp. 373-391). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-416008-8.00020-6

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., & Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134*(6), 1751-1764. https://doi.org/10.1093/brain/awr059

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J Pers Soc Psychol*, *85*(2), 348-362. https://doi.org/10.1037/0022-3514.85.2.348

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388. https://doi.org/https://doi.org/10.1016/0167-2681(82)90011-7

Hartley, C., & Sokol-Hessner, P. (2017). Affect is the foundation of value. In R. Davidson, A. Shackman, A. Fox, & R. Lapate (Eds.), *The nature of emotion*. Oxford University Press.

Hétu, S., Luo, Y., D'Ardenne, K., Lohrenz, T., & Montague, P. R. (2017a). Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. *Social cognitive and affective neuroscience*, nsx097-nsx097. https://doi.org/10.1093/scan/nsx097

Hétu, S., Luo, Y., D'Ardenne, K., Lohrenz, T., & Montague, P. R. (2017b). Human substantia nigra and ventral tegmental area involvement in computing social error signals during the ultimatum game. *Social Cognitive and Affective Neuroscience*, *12*(12), 1972–1982. https://doi.org/https://doi.org/10.1093/scan/nsx097

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91-106. https://doi.org/https://doi.org/10.1016/j.cognition.2017.03.006

Hooker, C., & Park, S. (2002). Emotion processing and its relationship to social functioning in schizophrenia patients. *Psychiatry Res*, *112*(1), 41-50. https://doi.org/https://doi.org/10.1016/S0165-1781(02)00177-4

Hutchinson, J. B., & Barrett, L. F. (2019). The Power of Predictions: An Emerging Paradigm for Psychological Research. *Current Directions in Psychological Science*, *28*(3), 280-291. https://doi.org/10.1177/0963721419831992

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods*, *48*(4), 1308-1317. https://doi.org/10.3758/s13428-015-0624-x

Ironside, M., Amemori, K.-i., McGrath, C. L., Pedersen, M. L., Kang, M. S., Amemori, S., Frank, M. J., Graybiel, A. M., & Pizzagalli, D. A. (2019). Approach-avoidance conflict in major depression: Congruent neural findings in human and non-human primates. *Biological psychiatry*. https://doi.org/https://doi.org/10.1016/j.biopsych.2019.08.022

Juechems, K., & Summerfield, C. (2019). Where Does Value Come From? *Trends in Cognitive Sciences*, *23*(10), 836-850. https://doi.org/https://doi.org/10.1016/j.tics.2019.07.012

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artif. Int. Res.*, *4*(1), 237–285.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Keren, H., O'Callaghan, G., Vidal-Ribas, P., Buzzell, G. A., Brotman, M. A., Leibenluft, E., Pan, P. M., Meffert, L., Kaiser, A., Wolke, S., Pine, D. S., & Stringaris, A. (2018). Reward Processing in Depression: A Conceptual and Meta-Analytic Review Across fMRI and EEG Studies. *Am J Psychiatry*, *175*(11), 1111-1120. https://doi.org/10.1176/appi.ajp.2018.17101124

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78-83. https://doi.org/10.1126/science.1108062

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange [10.1126/science.1108062]. *Science*, *308*(5718), 78. http://science.sciencemag.org/content/308/5718/78.abstract

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3771-3786. https://doi.org/10.1098/rstb.2008.0155

Kumar, P., Goer, F., Murray, L., Dillon, D. G., Beltzer, M. L., Cohen, A. L., Brooks, N. H., & Pizzagalli, D. A. (2018). Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology*, *43*(7), 1581-1588. https://doi.org/10.1038/s41386-018-0032-x

Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., & Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain*, *131*(8), 2084-2093. https://doi.org/10.1093/brain/awn136

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042-1060. https://doi.org/10.1037/a0020962

Lang, P. J., & Bradley, M. M. (2008). Appetitive and defensive motivation is the substrate of emotion. In *Handbook of approach and avoidance motivation.* (pp. 51-65). Psychology Press.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027-1038. https://doi.org/https://doi.org/10.1016/j.conb.2012.06.001

Lindquist, K. A., & Barrett, L. F. (2008). *Emotional complexity*. In *Handbook of emotions, 3rd ed.* The Guilford Press.

Liu, C., Chai, J. W., & Yu, R. (2016). Negative incidental emotions augment fairness sensitivity [Article]. *Scientific Reports*, *6*, 24892. https://doi.org/10.1038/srep24892

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In *Handbook of affective sciences.* (pp. 619-642). Oxford University Press.

Marin, R. S., Biedrzycki, R. C., & Firinciogullari, S. (1991). Reliability and validity of the Apathy Evaluation Scale. *Psychiatry Res*, *38*(2), 143-162.

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology*, *49*(4), 549-565. https://doi.org/10.1111/j.1469-8986.2011.01320.x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190. https://doi.org/https://doi.org/10.1016/j.jneumeth.2007.03.024

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332-345. https://doi.org/10.1037/0096-3445.128.3.332

Mellers, B. A., & McGraw, A. P. (2001). Anticipated Emotions as Guides to Choice. *Current Directions in Psychological Science*, *10*(6), 210-214. https://doi.org/10.1111/1467-8721.00151

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron*, *56*(1), 14-18.

Moutoussis, M., Rutledge, R. B., Prabhu, G., Hrynkiewicz, L., Lam, J., Ousdal, O.-T., Guitart-Masip, M., Fonagy, P., & Dolan, R. J. (2018). Neural activity and fundamental learning, motivated by monetary loss and reward, are intact in mild to moderate major depressive disorder. *PLoS One*, *13*(8), e0201451. https://doi.org/10.1371/journal.pone.0201451

Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in Cognitive Sciences*, *11*(11), 489-497. https://doi.org/https://doi.org/10.1016/j.tics.2007.08.013

Nielsen, L., Knutson, B., & Carstensen, L. L. (2008). Affect dynamics, affective forecasting, and aging. *Emotion (Washington, D.C.)*, *8*(3), 318-330. https://doi.org/10.1037/1528-3542.8.3.318

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex [Article]. *Nature Neuroscience*, *4*, 95. https://doi.org/https://doi.org/10.1038/82959

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042-1045. https://doi.org/10.1038/nature05051

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*(2), 148-158. https://doi.org/DOI 10.1038/nrn2317

Pessoa, L. (2009). How do emotion and motivation direct executive control? *Trends in Cognitive Sciences*, *13*(4), 160-166. https://doi.org/10.1016/j.tics.2009.01.006

Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annu. Rev. Psychol.*, *57*, 27-53.

Phelps, E. A., Lempert, K. M., & Sokol-Hessner, P. (2014). Emotion and Decision Making: Multiple Modulatory Neural Circuits. *Annual review of neuroscience*, *37*(1), 263-287. https://doi.org/10.1146/annurev-neuro-071013-014119

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715-734. https://doi.org/10.1017/S0954579405050340

Radloff, L. S. (1977). *The CES-D Scale: A self-report depression scale for research in the general population* [doi:10.1177/014662167700100306].

Richard, R., van der Pligt, J., & de Vries, N. (1996). Anticipated Affect and Behavioral Choice. *Basic and Applied Social Psychology*, *18*(2), 111-129. https://doi.org/10.1207/s15324834basp1802_1

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making [Review]. *Nat Rev Neurosci*, *15*(8), 549-562. https://doi.org/10.1038/nrn3776

Russell, J. A. (1980). A circumplex model of affect [doi:10.1037/h0077714]. *39*, 1161-1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion [doi:10.1037/0033-295X.110.1.145]. *Psychological Review*, *110*(1), 145-172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805-819. https://doi.org/10.1037/0022-3514.76.5.805

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493-502. https://doi.org/10.1037/0022-3514.57.3.493

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., & Dolan, R. J. (2017). Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. *JAMA Psychiatry*, *74*(8), 790-797. https://doi.org/10.1001/jamapsychiatry.2017.1713

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252. https://doi.org/10.1073/pnas.1407535111

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. *Science*, *318*(5850), 598-602. https://doi.org/10.1126/science.1142996

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*(5626), 1755-1758. https://doi.org/10.1126/science.1082976

Scherer, K. R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3459-3474. https://doi.org/10.1098/rstb.2009.0141

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward [10.1126/science.275.5306.1593]. *Science*, *275*(5306), 1593. http://science.sciencemag.org/content/275/5306/1593.abstract

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual review of neuroscience*, *23*, 473-500. https://doi.org/DOI 10.1146/annurev.neuro.23.1.473

Smith, D. V., Hayden, B. Y., Truong, T.-K., Song, A. W., Platt, M. L., & Huettel, S. A. (2010). Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. *The Journal of Neuroscience*, *30*(7), 2490. https://doi.org/10.1523/JNEUROSCI.3319-09.2010

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *Br J Psychiatry*, *167*(1), 99-103.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences*, *106*(13), 5035. https://doi.org/10.1073/pnas.0806761106

Sutton, R. S., & Barto, A. G. (1998). *Introduction to Reinforcement Learning*. MIT Press.

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*, *114*(23), 5982. https://doi.org/10.1073/pnas.1616056114

Treadway, M. T., & Zald, D. H. (2013). Parsing Anhedonia: Translational Models of Reward-Processing Deficits in Psychopathology. *Current Directions in Psychological Science*, *22*(3), 244-249. https://doi.org/10.1177/0963721412474460

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585-594. https://doi.org/10.1016/j.tics.2005.10.011

Whitton, A. E., Treadway, M. T., & Pizzagalli, D. A. (2015). Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current opinion in psychiatry*, *28*(1), 7-12. https://doi.org/10.1097/YCO.0000000000000122

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wulff, D. U., Haslbeck, J. M. B., Kieslich, P. J., Henninger, F., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: Detecting types in movement trajectories. In M. Schulte_Mecklenbeck, A. Kughberger, & J. G. Johnson (Eds.), *A Handbook of Process Tracing Methods*.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational Substrates of Norms and Their Violations during Social Exchange. *The Journal of Neuroscience*, *33*(3), 1099. https://doi.org/10.1523/JNEUROSCI.1642-12.2013

Supplementary information

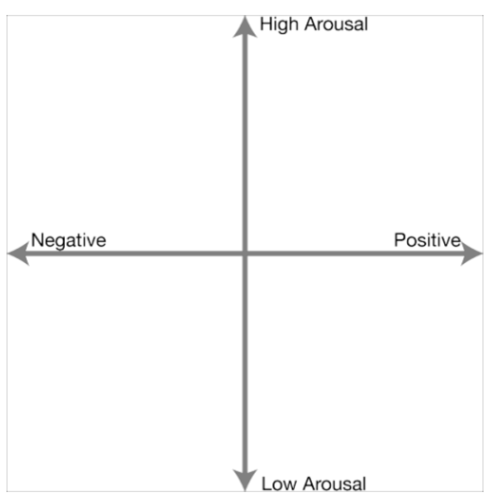
1 Supplementary Methods

1.1 Task Instructions

In Experiment 1, participants completed an emotion classification (EC) task before playing the Ultimatum Game (UG). In the EC, participants rated 20 feeling words on the *dynamic Affective Representation Mapping* (dARM) measurement, adapted from the affect grid used in past research (Russell et al., 1989). The 20 terms (*neutral, surprised, aroused, peppy, enthusiastic, happy, satisfied, relaxed, calm, sleepy, still, quiet, sluggish, sad, disappointed, disgusted, annoyed, angry, afraid, nervous*) were selected from past research to evenly represent the octants of the circumplex space (Feldman, 1995) or because they have previously been implicated in decisions to punish (Liu et al., 2016). Instructions from the emotion classification task were adapted from past literature (Russell et al., 1989).

<u>Instructions for the Emotion Classification Task</u>

The purpose of Part 1 is to study how people classify feelings. In this part, you will use diagrams like the one below to describe feelings. It is in the form of a square - a kind of map for feelings. The center of the square represents a neutral, average, everyday feeling. It is neither positive nor negative.



The right half of the grid represents pleasant feelings. The farther to the right the more pleasant. The left half represents unpleasant feelings. The farther to the left, the more unpleasant.

The vertical dimension of the map represents degree of arousal. Arousal has to do with how wide awake, alert, or activated a person feels - independent of whether the feeling is positive or negative. The top half is for feelings that are above average in arousal, and the lower half for feelings below average. The bottom represents sleep, and the higher you go, the more awake a person feels. At the top is maximum arousal. If you imagine a state we might call frantic excitement (remembering that it could be either positive or negative), then this feeling would define the top of the grid.

In this part of the study, you will be shown words that describe certain feelings. Try to think about what the feeling is like. Some people think about situations, or draw on memories of situations, that have made them feel that in the past. Please rate the feeling using the diagram above.

You can click on the diagram to rate the feeling any number of times, but be aware that the last click will be recorded as your final rating of that feeling.

Instructions for the Ultimatum Game

The purpose of Part 2 is to study how people make decisions.

*You will be making real decisions that affect the monetary outcomes of **YOURSELF** and **OTHERS**.*

You will be playing multiple rounds of a game with DIFFERENT partners every round.

Your partner has been allotted $1.00. You will be paired with a different partner each round and your partner has already decided how much of their $1 to offer you. For example, your partner can decide to:

1. keep $.95 and give $.05	*2. keep $.90 and give $.10*
3. keep $.85 and give $.15	*4. keep $.80 and give $.20*
5. keep $.75 and give $.25	*6. keep $.70 and give $.30*
7. keep $.65 and give $.35	*8. keep $.60 and give $.40*
9. keep $.55 and give $.45	*10. keep $.50 and give $.50*

IMPORTANT: Your partner can split their $1 in any amount as long as it is in $.05 increments.

After observing your partner make an offer, you will be asked to determine the final monetary outcome of both your partner and yourself. These options are labeled below:

1. ***Accept****: Agree to the proposed offer and keep both you and your partner's money the same.*
2. ***Reject****: Reject the offer and decrease both you and your partner's money to zero. Ultimately, you will decide how much money you and your partner actually receive.*

<u>EXAMPLE TRIAL</u>

During the task itself, you will be asked to make some predictions about the offer before receiving it and making a decision. First, you will be shown your partner's picture and name and asked how much of their $1.00 you think they will offer you. Here is an example of how this question will look:

After you answer, you will also be asked how you FEEL about the offer you expect to receive from your partner. Here is an example of how this question will look:

Once your partner has made his offer you will be asked how you FEEL about your partner's actual offer. Here is an example of how this question will look:

Lastly, you will be asked to decide whether to ***Accept*** *or* ***Reject*** *your partner's offer. Here is an example of how this question will look:*

In this example, you can see the options that will determine the monetary outcomes of both players.

Picking the ***Accept*** *option (on the left) will keep you and your partner's outcome the same as the offer.*
Picking the ***Reject*** *option (on the right) will reduce both you and your partner's outcome to zero.*

IMPORTANT: The choices that you will see will never be in the same order, so pay attention!

In short, you have the ability to determine the division of the monetary outcomes between your partner and yourself.

<u>BONUS PAY OUT</u>

At the end of the game, one trial from the task will be realized. That is, ALL Players will receive an additional BONUS for one of your decisions. This decision will be selected using a random number generator. If you selected the option on the left in the previous example, you would be paid out an additional $.20 and your partner would be paid out $.80. If you selected the option on the right, you would be paid out $0 and your partner would receive $0.

This bonus is in addition to the $2.50 you will make just for playing the game. You will be paid out this additional bonus within 3 days of finishing the game.

1.2 Experiment 1 – Ultimatum Game

Exclusion criteria. Participants were excluded from analysis if they failed to pass an attention check. In the instructions, we explicitly informed participants that "neutral" falls in the center of the dARM grid. We therefore excluded participants if they responded outside of a 100 x 100 pixel window in the center of the 500 x 500 pixel grid. Of the total of 398 participants we recruited on mTurk, 34 participants were excluded for failing this attention check, leaving N = 364 for the final analysis.

1.3 Experiment 2 – Replication

Methods. In Experiments 1, participants were asked for a reward prediction ("how much do you think [partner's name] will offer?"), followed by an emotion prediction ("how do you think you will feel about the offer?"). Our results showed that individuals relied more on emotion prediction errors than reward prediction errors when deciding whether to punish the proposer for their offer. It is possible that we obtained these results because the emotion prediction (which comes right after the reward prediction) saps the reward prediction of its saliency. If this were true, it would suggest that the introduction of an emotion probe artificially reduces the influence of reward prediction errors during choice and reward prediction errors may in fact be stronger than emotion prediction errors. To address this, we conducted a follow-up experiment which controlled for this potential methodological concern. Participants completed two blocks: 1) a reward-only block where they were only probed for reward predictions, and 2) a reward+emotion block replicating the design of all other experiments. All participants completed both blocks in counterbalanced order.

Exclusion criteria. Participants were excluded from analysis if they failed to pass the attention check described for Experiment 1. Of the 244 participants we recruited on Prolific, 16 participants were excluded for failing the attention check, resulting in a final participant sample of 228.

1.4 Experiment 3 – Justice Game

Exclusion criteria. Participants were excluded from analysis if they failed to pass the "neutral" rating attention check described for Study 1. Two of 75 in-lab participants were excluded for failing this attention check, resulting in a final sample of N = 73.

1.5 Experiment 4 – Mood disorders and Ultimatum Game

Methods. To explore the link between mood disorders and emotion and reward, we included the following questionnaires in our preregistered third study: the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977), the Temporal Experience of Pleasure Scale (TEPS) (Gard et al., 2006), the Behavioral Inhibition, Behavioral Activation Scales (BISBAS) (Carver & White, 1994), the Snaith-Hamilton Pleasure Scale (SHAPS) (Snaith et al., 1995), the Apathy Evaluation Scale (AES) (Marin et al., 1991), and the twenty-item Toronto Alexithymia Scale (TAS) (Bagby et al., 1994). Aside from the questionnaires, the experimental design was identical to Experiment 1, except that all participants played the Ultimatum Game as the Responder.

The main purpose of the study was to compare prediction error use in groups who present with significant levels of depression symptoms compared to those who do not. To further tease apart the relationship between reward and emotion, we included additional sub-questionnaires related to different facets of mood disorders, including pleasure, anhedonia, avolition, and reward. The

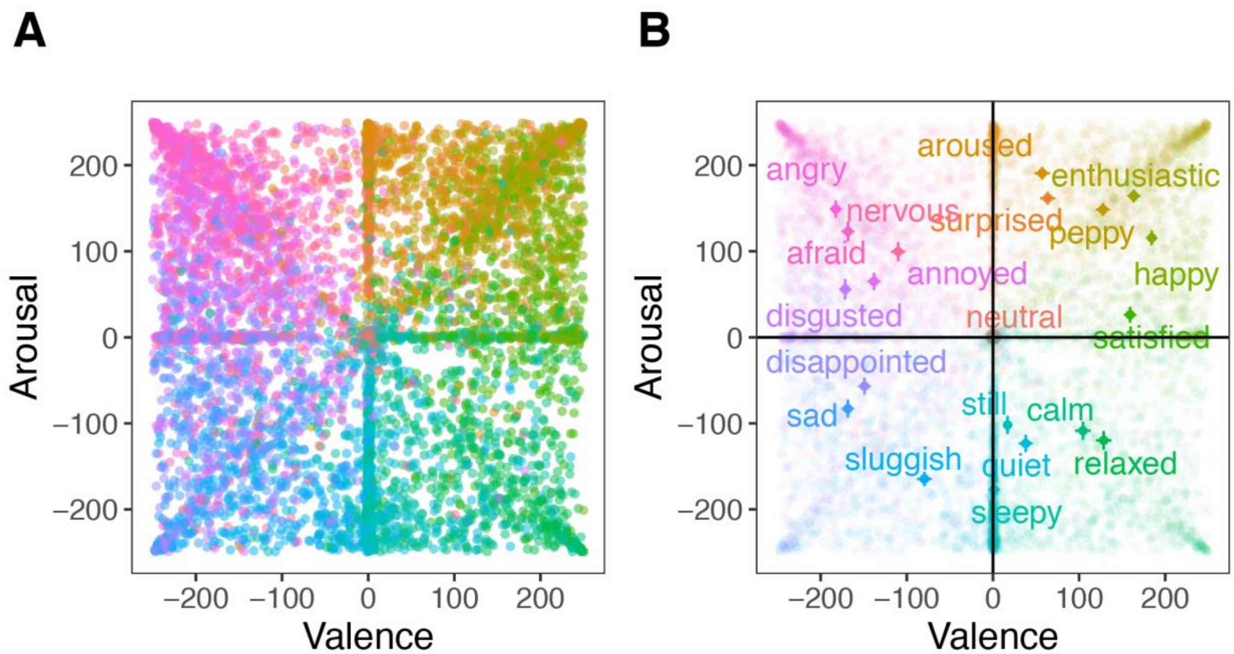
CES-D was the most direct measure of a mental health disorder that implicates impairments in both reward and emotion. However, we also analyzed the results of the other questionnaires, which provide additional and converging evidence about the roles of emotion and reward prediction errors during social choice.

Exclusion criteria. Participants were excluded from analysis if they failed to pass the attention check described for Experiment 1. Of the 508 participants we recruited on mTurk, 157 participants were excluded for failing this attention check, resulting in a final participant sample of N = 351.

2 Supplementary Results

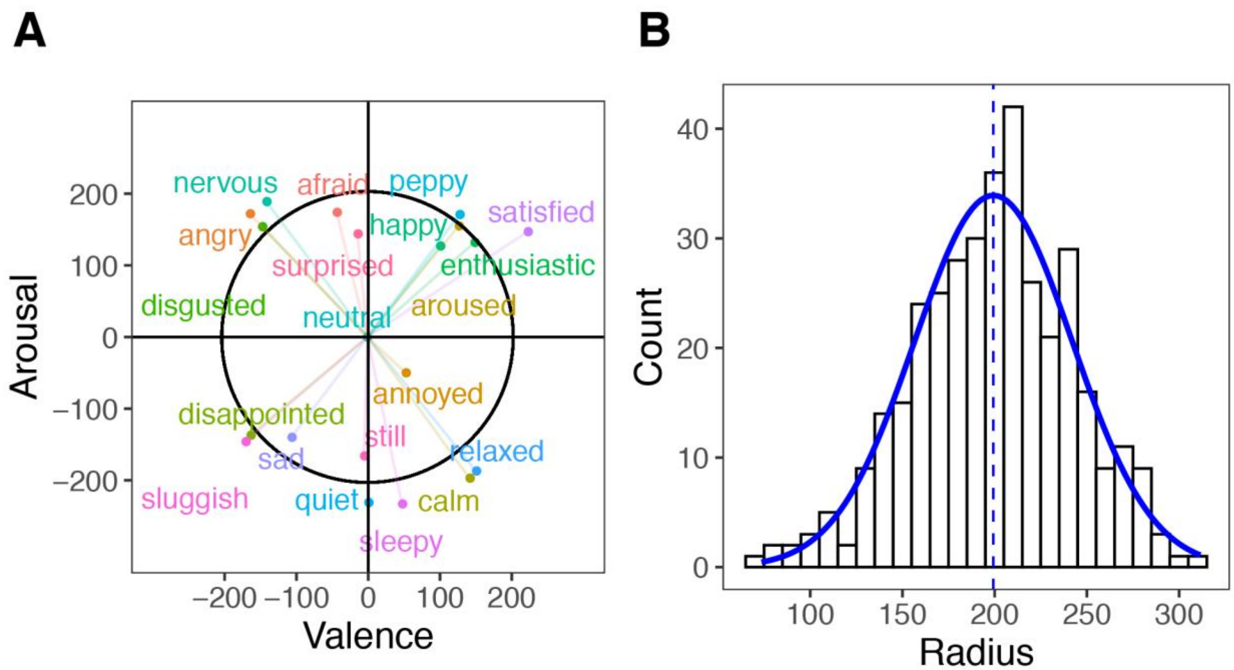
2.1 Emotion Classification Results

Circumplex structure of affect. Decades of research have shown that general semantic knowledge about valence and arousal can be represented in a circumplex structure (Russell, 1980). Results from our adapted measure replicate essential findings from prior research. First, the individual ratings of each feeling cluster along the octants (i.e., the axes and diagonal) of the circumplex space (Supplementary Figure 1A). Second, the average ratings of the feelings form a circular structure characteristic of the affective circumplex found in past work (Supplementary Figure 1B).



Supplementary Figure 1. Emotion Classification Task. A) All feeling ratings. Individual ratings of each feeling label are shown in the dARM space and are color-coded by their feeling label. Each dot reflects one individual's data per feeling word. **B) Average feeling ratings**. The average ratings of the 20 feeling labels, overlaid on top of individual ratings. Error bars represent 95% CIs.

Individual variation in emotional range. While the overall circular representation was highly consistent across participants, there were also considerable individual differences in how participants rated emotions. To further probe these differences, we constructed a circumplex for each participant by calculating the average distance of each feeling label compared to "neutral" (the center of the participants' circumplex; Supplementary Figure 2A). The radius of the resulting structure represents participants' range of emotional experiences within this emotion space. Individuals with greater radiuses can therefore be thought of as showing higher emotional range (Supplementary Figure 2B).



Supplementary Figure 2. A) Example participant's circumplex. The radius of the circumplex is calculated by taking the average distance of a participant's feeling ratings from their own rating of 'neutral', thus representing the range of their emotional representations. **B) Emotional range histogram**. The histogram displays the distribution of participants' emotional range. Larger values indicate greater emotion range and a stronger distinction between individual feeling labels. For example, similar discrete emotions such as 'peppy' and 'happy' would likely be represented farther apart in someone with a high emotional range. The solid blue line overlays the normal distribution for this histogram, and the dashed blue line indicates the mean of the distribution.

2.2 Experiment 1 - Ultimatum Game Results

Responders and third-party deciders respond to unfairness in similar ways. Participants either played the UG as the Responder or a non-vested third party. Using a logistic mixed-effects model, we examined whether the effect of unfairness on rejecting the offer depended on the participants' role. Results showed a main effect of role, such that third parties punished more than Responders, but no significant interaction between unfairness and role (Supplementary Table 1). Thus, despite the difference in their roles, all participants responded similarly to unfair offers. We then examined whether the use of any prediction error (reward, valence, and arousal) differed among third parties and Responders. Results showed no significant interactions between emotion prediction errors and role (Supplementary Table 2), but there was a trending interaction showing that third parties used reward prediction errors less than Responders ($p = .08$). Because third parties and Responders behaved similarly in our task, we collapsed the groups for all reported analyses in Experiment 1. To ensure that our effects are not driven by participants in a single condition, we analyzed each role separately and found that valence prediction errors are significantly more predictive than reward prediction errors for both third parties and Responders.

Experiment 1: Third-parties and responders treat unfairness similarly

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-3.51	0.55	-6.41	**<0.001**
Unfairness	6.51	0.48	13.61	**<0.001**
Role [Responder]	-1.97	0.74	-2.65	**0.008**
Unfairness:Role [Responder]	-0.68	0.56	-1.23	0.219
N_{sub}	364			
Observations	7280			

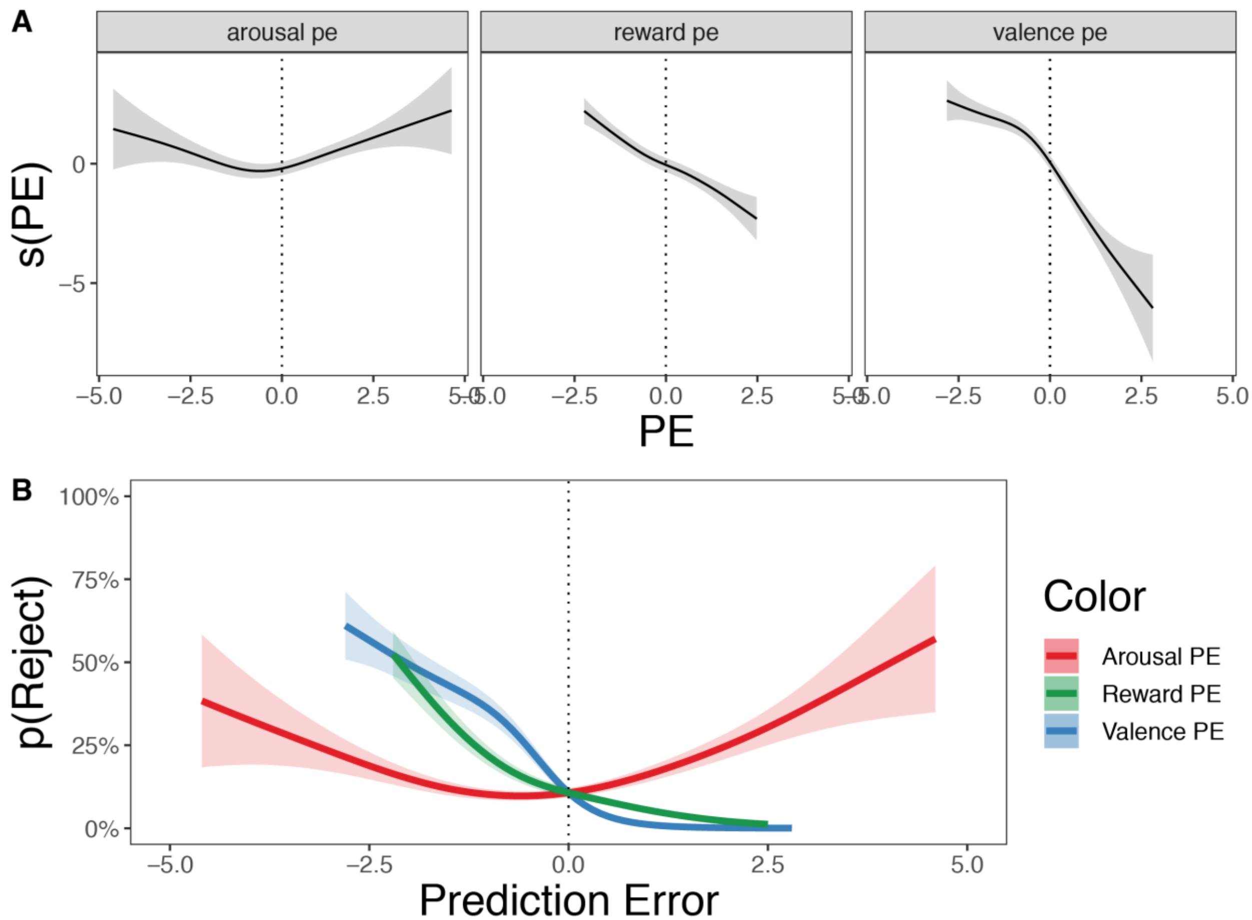
Supplementary Table 1. *Unfairness is defined as the amount of money kept by Player A and was normalized (scaled and mean-centered). Role is a binary variable with third-party (0) and responder (1). The model includes subject-specific random intercepts and slopes for Unfairness.*

Experiment 1: Third-parties and responders treat prediction errors similarly

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.02	0.23	-8.73	**<0.001**
Valence PE	-1.97	0.22	-9.09	**<0.001**
Role [Responder]	-1.08	0.32	-3.33	**0.001**
Arousal PE	0.55	0.13	4.32	**<0.001**
Reward PE	-0.87	0.18	-4.70	**<0.001**
Valence PE:Role [Responder]	0.08	0.30	0.26	0.793
Arousal PE:Role [Responder]	0.00	0.17	0.02	0.982
Reward PE:Role [Responder]	-0.47	0.26	-1.77	0.076
N_{sub}	364			
Observations	7280			

Supplementary Table 2. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Role is a binary variable with third-party (0) and responder (1). The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs.*

Nonparametric analysis of emotion and reward prediction errors. As suggested by Kahneman & Tversky's classic work on Prospect Theory (Kahneman & Tversky, 1979), it is possible that reward has a nonlinear effect on decision-making. We therefore used nonparametric regression analysis to obtain data-driven model predictions relating prediction errors to choices. Specifically, we used a generalized additive mixed effect model (GAMM), which models choices as a summation of nonlinear transformations for each of the three prediction errors (Wood, 2011). Using this approach, we replicate the results of the additive logistic model from the main text (Supplementary Figure 3A). The marginal contribution of each prediction error (i.e., each semi-partial estimate accounting for unique variance) shows that valence prediction errors have a stronger unique contribution to choice than reward prediction errors (Supplementary Table 3; Supplementary Figure 3B).



Supplementary Figure 3. GAMM analysis of prediction errors. A) Smooth spline transformations. *Plots show the transformation of the linear prediction error into a nonlinear smooth spline form.* ***B) Predictive plot for additive GAMM model.*** *Plot shows the model probability of rejecting the offer given the unique contributions of each prediction error modelled nonlinearly with a smoothing spline. Error bars are* $\pm$ *1 SE.*

Experiment 1: Non-parametric regression of prediction errors

	Estimates		
Predictors	*Effective DF*	*F*	*p*
s(Valence PE)	4.22	87.33	**<0.001**
s(Arousal PE)	3.75	14.59	**<0.001**
s(Reward PE)	2.87	46.30	**<0.001**

Supplementary Table 3. *GAMMs allow for nonparametric relationships by modeling each prediction error as a smooth penalized regression spline. Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. Before smoothing splines were applied, prediction errors were scaled but not mean-centered. The model includes subject-specific random intercepts.*

Interactions between prediction errors and their relationship with punitive decisions. We performed an exploratory logistic regression analysis modeling the three-way interaction between all prediction errors and every possible two-way interaction between the prediction errors. Results reveal a significant three-way interaction between arousal, valence, and reward prediction errors (Supplementary Table 4), indicating that the effect of each prediction error is conditional on the values of the other two prediction errors. Examining the two-way interactions shows a significant interaction between valence and arousal prediction errors, reward and arousal prediction errors, but not reward and valence prediction errors. This suggests that valence and reward prediction errors may be dependent on arousal prediction errors in determining choices, but not on each other. Likelihood ratio tests between the interactive prediction error model and the additive prediction error model show that the interactive model fit better ($\chi^2(4) = 40.00$, $p < .001$).

Experiment 1: Interactions between prediction errors

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.50	0.18	-13.99	**<0.001**
Valence PE	-1.87	0.17	-10.85	**<0.001**
Arousal PE	0.44	0.13	3.55	**<0.001**
Reward PE	-1.09	0.16	-6.97	**<0.001**
Valence PE:Arousal PE	-0.36	0.11	-3.22	**0.001**
Valence PE:Reward PE	-0.05	0.10	-0.52	0.604
Arousal PE:Reward PE	-0.30	0.11	-2.65	**0.008**
Valence PE:Arousal PE:Reward PE	-0.37	0.07	-4.91	**<0.001**
N_{sub}	364			
Observations	7280			

Supplementary Table 4. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Role is a binary variable with third-party (0) and responder (1). The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs.*

Expectations and prediction errors independently predict decisions to punish. We tested whether prediction errors play an independent role from expectations in predicting decisions to punish. We modeled our analysis after prior work where expectations were included with prediction errors in the same model (Rutledge et al., 2014). Results show that valence, and reward PEs still independently explain significant variance in decisions to punish even after controlling for the contributions of expectations (Supplementary Table 5).

Experiment 1: Expectations and prediction errors independently explain punitive choices

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	1.25	0.45	2.80	**0.005**
Valence Expectation	-2.26	0.19	-12.11	**<0.001**
Arousal Expectation	-0.10	0.13	-0.80	0.424
Reward Expectation	-6.92	0.46	-14.91	**<0.001**
Valence PE	-2.97	0.26	-11.47	**<0.001**
Arousal PE	0.01	0.13	0.06	0.955
Reward PE	-4.74	0.31	-15.33	**<0.001**
N_{sub}	364			
Observations	7280			

Supplementary Table 5. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 points are meaningful. The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs.*

Experienced reward is a stronger predictor of punishment decisions than experienced emotions. Although the primary focus of our present research was to understand the role of reward and emotion expectations in social choice, we can also contextualize our findings by limiting our analysis to participants' experiences during our task (i.e., ignoring expectations), which emulates past approaches to studying emotion and social choice. Using a logistic mixed-effects model, we directly compared the relative contributions of the experienced reward (i.e., amount offered) and experienced emotions (i.e., valence and arousal ratings on the circumplex) upon receiving the offer. Results show that both the experience of valence and reward, but not arousal, predict punishment (Supplementary Table 6).

Experiment 1: Experienced reward predicts decisions to punish better than experienced valence or arousal

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-5.19	0.42	-12.23	**<0.001**
Experienced Valence	-2.57	0.19	-13.78	**<0.001**
Experienced Arousal	-0.10	0.12	-0.82	0.412
Experienced Reward	-4.31	0.35	-12.25	**<0.001**
N_{sub}	364			
Observations	7280			

Supplementary Table 6. *Experienced Reward is defined as the amount of money offered by Player A and was normalized (scaled and mean-centered). Experienced Valence and Arousal are the experienced emotions on each dimension when receiving the offer and were scaled but not mean-centered (as the 0 point on each scale represents a neutral feeling). The model includes subject-specific random intercepts and slopes for Experienced Reward.*

Expectation and Prediction Error Model Comparisons. To better understand how expectations and prediction errors contributed to choice, we compared models which contained expectation and PE terms for reward only (m4a), reward and valence (m4b), and reward, valence, and arousal (m4c). Likelihood Ratio Tests reveal that a model which only includes reward expectations and reward PEs (m4a) fit the data worse than a model which added valence expectations and valence PEs (m4b; $\chi^2(2) = 338.39,\ p < .001$). However, adding arousal expectations and arousal PEs (m4c) did not improve the model fit compared to m4b ($\chi^2(2) = 1.428,\ p = .49$), as they had comparable BICs (m4b: 3206.1, m4c: 3222.5).

Controlling for Expectations in the Interactive Model. We used nested mixed-effects regressions to test whether the interactions between prediction errors found in Supplementary Table 4 still hold when controlling for reward, valence, and arousal expectations. We compared two models: one which included all PEs and their interactions (the “interactive” model), and one which adds expectations (reward, valence, and arousal) on top of the interactive model. Likelihood ratio tests from Experiment 1 reveals that adding expectations improves the fit of the data $\chi^2(3) = 1847.9$, $p < .001$. In this model, both valence and reward PEs remain significant after accounting for expectations, although arousal PEs no longer contribute significant explanatory variance. Furthermore, the two interactions previously found in Supplementary Table 4 are weakened to be only marginally significant (Supplementary Table 7).

Experiment 1: Controlling for Expectations in the Interactive Model

	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	1.09	0.43	2.51	**0.012**
Valence Expectations	-2.27	0.18	-12.75	**<0.001**
Arousal Expectations	-0.11	0.12	-0.88	0.380
Reward Expectations	-6.73	0.45	-15.12	**<0.001**
Valence PE	-3.16	0.20	-15.42	**<0.001**
Arousal PE	-0.09	0.13	-0.70	0.486
Reward PE	-4.54	0.30	-15.31	**<0.001**
Valence PE: Arousal PE	0.22	0.12	1.79	0.073
Valence PE: Reward PE	-0.07	0.12	-0.59	0.559
Arousal PE: Reward PE	-0.23	0.12	-1.84	0.065
Valence PE: Arousal PE: Reward PE	0.02	0.08	0.26	0.792
N_{sub}	364			
Observations	7280			

Supplementary Table 7. *All variables were scaled but not mean-centered, as the 0 points are meaningful. The model includes subject-specific random intercepts and random slopes for Reward PEs.*

Computational Modeling Results for the Ultimatum Game. We used more formal modeling techniques to test whether emotion prediction errors are merely capturing variance which would normally be soaked up by modeling individual differences in the subjective valuation of reward. We drew from past research to create a computational utility model with an exponential scaling parameter (Kahneman & Tversky, 1979; Rutledge et al., 2014; Sokol-Hessner et al., 2009), which captures individual differences in how participants (nonlinearly) value rewards of varying magnitudes. We used the following equations to transform objective reward magnitudes into subjective value before calculating the resulting subjective reward prediction error (sRPE):

$$sRPE = reward_{actual}^{\lambda} - reward_{prediction}^{\lambda}$$
$$where\ 0 \le \lambda \le 1$$

By bounding lambda, this utility model captures the diminishing marginal utility of reward, in the same way as many traditional accounts, including Prospect Theory (Kahneman & Tversky, 1979). We incorporated subjective reward prediction errors into a utility model by adding an additional free parameter w_1, which specifies the subject's weight on model-derived sRPEs (i.e., how important sRPEs are in predicting choices):

$$m2a\ model:$$
$$utility_{accept} = w_1 * sRPE$$
$$utility_{reject} = 0$$

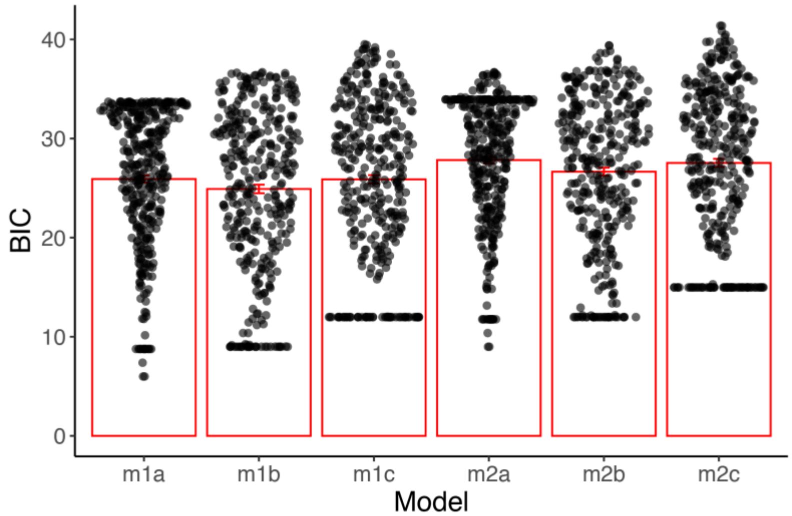
Standard models of the Ultimatum Game place the utility of rejecting at 0 (because both players receive no money), allowing PEs to modify the relative utility of accepting (Camerer, 2003). The choice rule was computed by placing the utility values for each decision into a softmax function:

$$p(accept) = \frac{e^{utility_{accept}* \beta}}{e^{utility_{accept}* \beta} + e^{utility_{reject}* \beta}}$$

$$p(reject) = 1 - p(accept)$$

To be thorough, we compared model likelihoods for this model (*m2a*) with two additional models, which sequentially added valence PEs *(m2b:* $utility_{accept} = w_1 * sRPE + w_2 * VPE$ *)* and arousal PEs *(m2c:* $utility_{accept} = w_1 * sRPE + w_2 * VPE + w_3 * APE$*)* into the utility function. Finally, we implemented a computational analogue of the mixed-effects regression reported in the manuscript, which used the original formulation of empirical reward prediction errors *(m1a:* $utility_{accept} = w_1 * RPE$, *m1b:* $utility_{accept} = w_1 * RPE + w_2 * VPE$, *m1c:* $utility_{accept} = w_1 * RPE + w_2 * VPE + w_3 * APE$*)*. Each component of the prediction error (expectations and experience) was normalized to the range [0, 1] before prediction errors were calculated (experience – expectation) to make the weights comparable. Models labeled with "*m1*" use empirical reward prediction errors (defined as monetary reward, minus monetary reward prediction), while models labeled with "*m2*" use subjective reward prediction errors (defined above).

Results show that the subjective reward PE model (*m2a*)—which incorporates the non-linear valuation of reward into the prediction error itself—provides a worse fit than the model that uses empirical reward PEs (*m1a*; Supplementary Figure 4; Supplementary Table 8; *Model 2a vs 1a:* $t(363) = 32.06$, $p < .001$). Although both model fits are improved by adding valence PEs to models with only reward PEs (*Model 1b vs 1a:* $t(363) = -4.00$, $p < .001$; *Model 2b vs 2a:* $t(363) = -4.58$, $p < .001$)), we again find that the model using empirical RPEs (*m1b*) outperforms the model using subjective RPEs (*m2b*; $t(363) = -15.27$, $p < .001$). Finally, we find that adding arousal in addition to reward and valence PEs does not provide a better fit and models omitting arousal PEs fit best (*Model 1c vs 1b:* $t(363) = 5.19$, $p < .001$; *Model 2c vs 2b:* $t(363) = 4.52$, $p < .001$). Comparing individual fits for m1b and m1c revealed that 19.5% of participants were better fit by m1c, highlighting the potential variability in use of arousal PEs. Finally, again a model with empirical RPEs (*m1c*) out performs the corresponding model using subjective RPEs (*m2c;* $t(363) = -12.07$, $p < .001$).

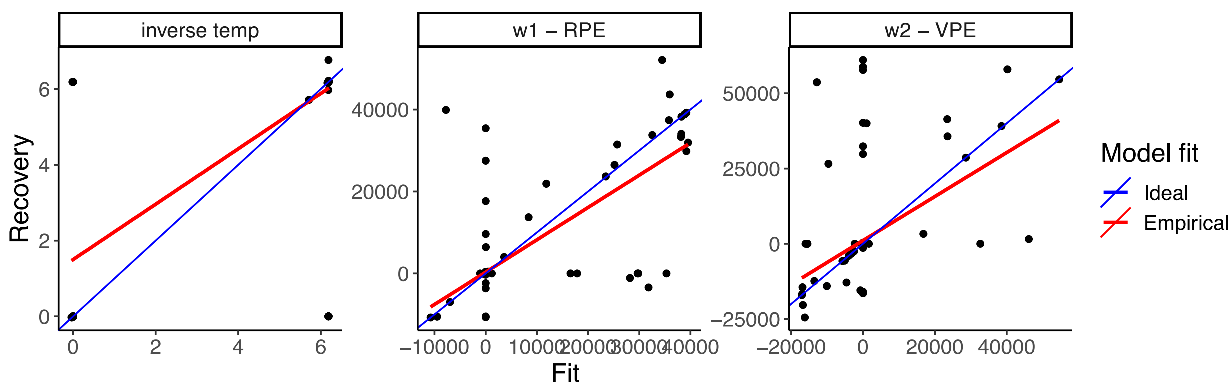


Supplementary Figure 4. Model comparisons for different PE models. *The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Individual points represent subject-level BIC values, while the red bars represent the group average.*

Model	Parameters	Model BIC (± SEM)	BIC – BIC_{model_1b}
1a	2	*25.91 (± 0.37)*	1.00
1b	3	*24.91 (± 0.43)*	*N/A*
1c	4	*25.88 (± 0.43)*	*0.97*
2a	3	*27.83 (± 0.34)*	*2.92*
2b	4	*26.65 (± 0.40)*	*1.74*
2c	5	*27.54 (± 0.41)*	*2.63*

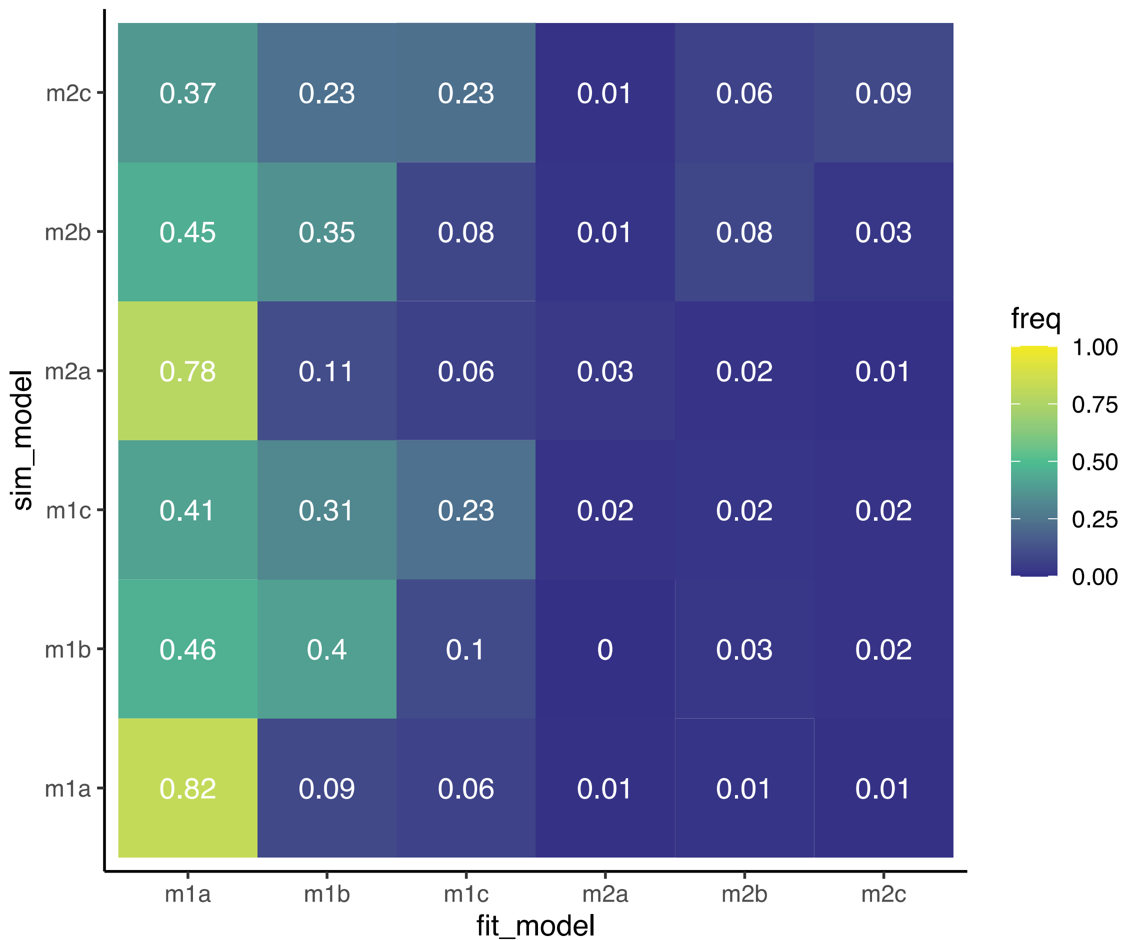
Supplementary Table 8. Model comparisons for different PE models*. The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Model BIC shows the average BIC for the model alongside the standard error of the mean.*

Parameter Recovery. We performed parameter recovery on the winning model (*m1b*) by first identifying the best-fitting parameters for that model, which include weights on each prediction error (w1 and w2 for reward and valence PEs, respectively), as well as an inverse temperature term (beta) in the softmax choice rule. Once we identified the best-fitting parameters, we simulated data with these known parameter values, and re-fit the model on the simulated data (Nassar, Waltz, Albrecht, Gold, & Frank, 2021). When comparing the recovered parameters to their true values, we found evidence for successful parameter recovery using our model (Figure S5). The Spearman correlations between the fitted and recovered parameters all show that information is being successfully captured by the model (inverse temp r = .58; w1/rPE r = .76; w2/vPE r = .76). The strength of these correlations shows we can recover this model's parameters despite the relatively low number of trials (20) per participant (Supplementary Figure 5).



Supplementary Figure 5. Parameter recovery of m1b. *Black dots represent individual subjects while the blue and red lines represent the ideal and empirical model fits. Each panel shows the recovery for one of the four parameters included in the model.*

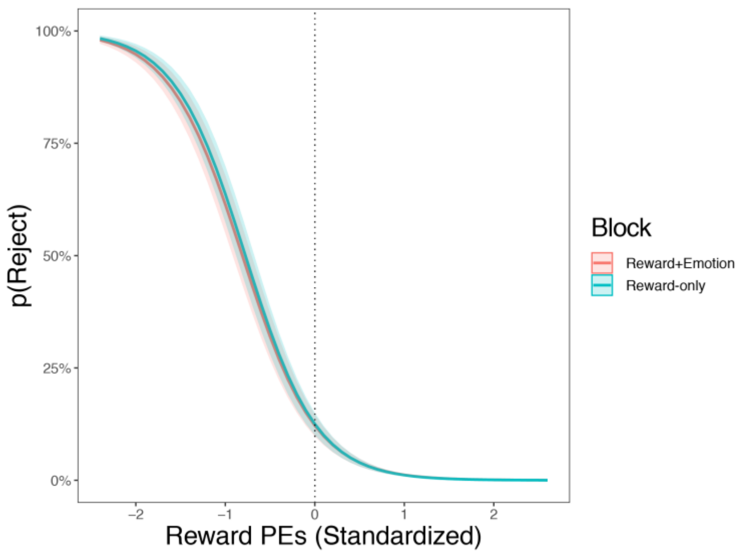
We also examined model identifiability for each of the six models by simulating data using subject-level PEs and the best fitting parameters for each model. The simulated data was then fit to each model separately, and the best fitting model was assessed with BIC. The results are summarized below in the confusion matrix (Supplementary Figure 6), where the numbers correspond to the percentage of simulations for which the model from the empirical data (x-axis) best fit the simulated model's data (y-axis). Perfect recovery would yield 0 everywhere but along the diagonal (which should have a value of 100)—although in practice, this is rarely the case (e.g., Wilson & Niv, 2012). Results show that model 1a (rPE only) is highly recoverable, though there is some confusion between this model and model 1b (rPE + vPE) and 1c (rPE, vPE, and aPE). We interpret this as providing even stronger evidence for our hypothesis. Even when the data was generated by m1b or m1c (which incorporates emotion prediction errors), there is a bias for the model-fitting process to favor m1a (which only incorporates reward PEs). Despite this bias, our model comparison still reveals that m1c (incorporating all three PEs) provides the best fit to participants' data, implying that our effect is robust enough to overcome the bias in the model-fitting process.



Supplementary Figure 6. Model Identifiability. *Confusion matrix for model recovery in the UG. Choices were simulated according to each model and then re-fit according to all six models. Darker colors indicate a lower probability that the fit model is the best model for the simulated data. Lighter colors represent cases with high probabilities.*

2.3 Experiment 2 - Replication Results

Predictive strength of reward PEs does not differ across block types. The most straightforward analysis is to compare the predictive strength of reward prediction errors across the two block types (reward-only vs. reward+emotion) using a mixed-effects logistic regression. Results reveal that the predictive power of reward prediction errors does not significantly depend on whether participants are also reporting the predictions and experiences of emotion (Supplementary Table 9; Supplementary Figure 7). That is, reward prediction errors are not robbed of their saliency when emotion prediction errors are measured in tandem with reward predictions.



Supplementary Figure 7. Reward prediction errors are not modulated by block type. *The strength of reward prediction errors predicting choices to reject are shown for the reward+emotion block (red) and reward-only block (blue). Error bars reflect ± 1 SE.*

Experiment 2: RPEs do not differ by block

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-1.94	0.22	-8.66	**<0.001**
Reward PE	-2.51	0.13	-19.36	**<0.001**
Block	-0.03	0.07	-0.42	0.677
Reward PE:Block	0.07	0.09	0.75	0.455
N_{sub}	228			
Observations	9120			

Supplementary Table 9. *Reward PEs are calculated as the difference between the experienced and predicted reward. Reward PEs were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Block is a binary variable with reward-only block (0) and reward+emotion block (1). The model includes subject-specific random intercepts and random slopes for Reward PEs.*

Reward predictions error use does not habituate over time. In our preregistration, we specified that we would examine potential habituation effects. Our intuition was that reward prediction errors might have a weaker effect in the second block than the first due to participants habituating to the task. In our model, this effect corresponds to the interaction between reward prediction errors and the block number (i.e., whether the data comes from the first versus second block, regardless of the type). Our results reveal that reward prediction errors actually exert a stronger influence over choice in the second block compared to the first block (Supplementary Table 10). This suggests the opposite of a habituation effect (i.e., a sensitization effect), where participants' reliance on reward prediction errors increases over time.

Experiment 2: Reliance on Reward PEs increases over time

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.03	0.23	-8.94	**<0.001**
Reward PE	-2.32	0.13	-18.24	**<0.001**
Block Order [second]	0.11	0.07	1.46	0.144
Reward PE:Block Order [second]	-0.35	0.09	-3.78	**<0.001**
N_{sub}	228			
Observations	9120			

Supplementary Table 10. *Reward PEs are calculated as the difference between the experienced and predicted reward. Reward PEs were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Block order is a binary variable with first block (0) and second block (1). The model includes subject-specific random intercepts and random slopes for Reward PEs.*

Emotion probes do not change reward prediction error strength. To be exhaustive in our investigation, there is one additional analysis we can conduct. We can compare the effects of reward prediction errors made during only the first block of the task for all participants. This is a between-subjects analysis because half of participants received the reward+emotion block first while the other half received the reward-only block first. If the aforementioned methodological concern holds true, then participants who completed the reward+emotion block first should have weaker or less predictive reward prediction errors, compared to those who completed the reward-only block first. Instead, results show no significant differences between participants' use of reward PEs: participants who reported both their emotion and reward predictions did not use reward PEs any differently than participants who reported their reward predictions only (Supplementary Table 11). Thus, in situations where participants are not asked to make an emotional prediction, it does not appear that they rely any more on reward prediction errors. This indicates that the additional emotion prediction probe is not altering how the participant treats the reward prediction error. These additional data provide strong evidence that asking participants to introspect about their emotions does not change how much they use reward prediction errors.

Experiment 2: No difference in Reward PEs across conditions in the first block

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-1.83	0.28	-6.54	**<0.001**
Reward PE	-2.20	0.16	-13.37	**<0.001**
Condition	0.07	0.38	0.17	0.863
Reward PE:Condition	-0.08	0.21	-0.39	0.696
N_{sub}	228			
Observations	4560			

Supplementary Table 11. *Reward PEs are calculated as the difference between the experienced and predicted reward. Reward PEs were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Condition is a binary variable which refers to which block participants received first with reward-only (0) and reward+emotion (1). Condition was counterbalanced across participants, and the data only includes the first block participants completed. The model includes subject-specific random intercepts and random slopes for Reward PEs.*

2.4 Experiment 3 – Justice Game Results

Emotion prediction errors predicts decisions to punish and reverse. Three trials were excluded from the Accept/Punish choice set and one trial was excluded from the Compensate/Reverse choice set due to a technical error where reward predictions were recorded as less than 0 (i.e., negative money). We used logistic mixed-effects regressions to examine how emotion prediction errors predict punitive decisions in the Accept/Punish choice set (Supplementary Table 12) and Compensate/Reverse choice set (Supplementary Table 13).

Experiment 3: Valence and reward prediction errors predict decisions to Punish/Accept

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-0.24	0.33	-0.73	0.467
Valence PE	-1.26	0.24	-5.18	**<0.001**
Arousal PE	0.04	0.18	0.21	0.830
Reward PE	-1.06	0.29	-3.60	**<0.001**
N_{sub}	73			
Observations	639			

Supplementary Table 12. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 points are meaningful. The dependent variable, choice, is a binary variable with accept (0) and punish (1). The model includes subject-specific random intercepts and random slopes for Reward PEs. Data only includes decisions to Punish versus Accept.*

Experiment 3: Valence, and reward prediction errors predict decisions to REVERSE/COMPENSATE

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-1.69	0.32	-5.36	**<0.001**
Valence PE	-0.97	0.26	-3.74	**<0.001**
Arousal PE	0.18	0.16	1.11	0.265
Reward PE	-0.79	0.32	-2.48	**0.013**
N $_{sub}$	73			
Observations	641			

Supplementary Table 13. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 points are meaningful. The dependent variable, choice, is a binary variable with compensate (0) and reverse (1). The model includes subject-specific random intercepts and random slopes for Reward PEs. Data only includes decisions to Reverse versus Compensate.*

Expectation and Prediction Error Model Comparisons. To better understand how expectations and prediction errors contributed to choice, we compared models which contained expectation and PE terms for reward only (m4a), reward and valence (m4b), and reward, valence, and arousal (m4c) collapsing across trials which contained decisions to Compensate or Reverse and Accept or Punish for power. Likelihood Ratio Tests reveal that a model which only includes reward expectations and reward PEs (m4a) fit the data worse than a model which added valence expectations and valence PEs (m4b; $\chi^2(2) = 67.31,\ p < .001$). However, adding arousal expectations and arousal PEs (m4c) did not improve the model fit compared to m4b ($\chi^2(2) = 0.14,\ p = .93$), as they had comparable BICs (m4b: 1186.2, m4c: 1200.4).

Additive and Interactive Model Comparisons. We compared models which contained all three prediction errors additively to an interactive model. Likelihood Ratio Tests reveal that a model which includes interactions between prediction errors fits better than a model with no interactions ($\chi^2(4) = 20.89,\ p < .001$).

Controlling for Expectations in the Interactive Model. We used nested mixed-effects regressions to test whether the interactions between prediction errors still hold when controlling for reward, valence, and arousal expectations. We compared two models: one which included all PEs and their interactions (the "interactive" model), and one which adds expectations (reward, valence, and arousal) on top of the interactive model. Likelihood ratio tests from Experiment 3 reveals that adding expectations improves the fit of the data $\chi^2(3) = 94.86,\ p < .001$. In this model, both valence and reward PEs remain significant after accounting for expectations, although arousal PEs no longer contribute significant explanatory variance (Supplementary Table 14).

Experiment 3: Controlling for Expectations in the Interactive Model

	Estimates			
Predictors	*Log-Odds*	*std. Error*	*Statistic*	*p*
Intercept	0.05	0.41	0.11	0.911
Valence Expectations	-0.74	0.15	-4.94	**<0.001**
Arousal Expectations	0.00	0.11	0.01	0.990
Reward Expectations	-1.18	0.37	-3.23	**0.001**
Valence PE	-1.21	0.20	-6.03	**<0.001**
Arousal PE	-0.09	0.17	-0.53	0.594
Reward PE	-0.85	0.23	-3.65	**<0.001**
Valence PE: Arousal PE	-0.33	0.18	-1.84	0.066
Valence PE: Reward PE	0.04	0.12	0.32	0.749
Arousal PE: Reward PE	0.08	0.18	0.45	0.653
Valence PE: Arousal PE: Reward PE	-0.14	0.12	-1.18	0.240
N_{sub}	73			
Observations	1288			

Supplementary Table 14. *All variables were scaled but not mean-centered, as the 0 points are meaningful. The model includes subject-specific random intercepts and random slopes for Reward PEs.*

Computational Modeling of Justice Game. We performed the same computational modeling analysis applied in Experiment 1 to Experiment 3. Once again, our computational utility models compare different combinations of PE types to predict choice. For statistical power, we collapse across trials that include decisions to Accept or Punish the offer and choices to Compensate or Reverse, resulting in a total of 18 trials per subject. As we did with Experiment 1, we implemented a computational analogue of the mixed-effects regression reported in the manuscript, which used the original formulation of empirical reward PEs:

$$m1a\text{: } utility_{accept} = w_1 RPE \tag{2a}$$
$$m1b\text{: } utility_{accept} = w_1 RPE + w_2 VPE \tag{2b}$$
$$m1c\text{: } utility_{accept} = w_1 RPE + w_2 VPE + w_3 APE \tag{2c}$$

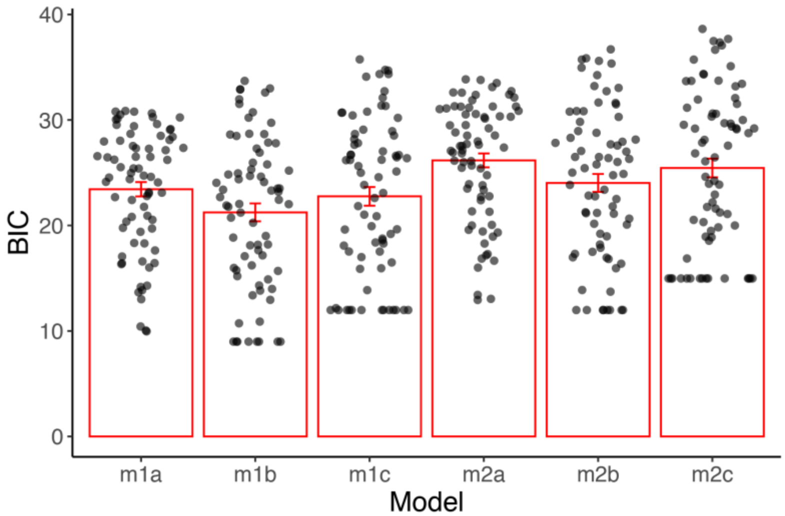
We also implemented models which incorporate individual differences in how participants (nonlinearly) value rewards of varying magnitudes. We used the following equations to transform objective reward magnitudes into subjective value before calculating the resulting subjective reward prediction error (sRPE):

$$\begin{gathered} sRPE = reward_{actual}^{\lambda} - reward_{prediction}^{\lambda} \\ where\ 0 \leq \lambda \leq 1 \end{gathered} \tag{3}$$

Models prefixed with "m2" refer to models which use sRPE in place of empirical RPEs, and are defined below:

$$m2a\text{: } utility_{accept} = w_1 sRPE \tag{4a}$$
$$m2b\text{: } utility_{accept} = w_1 sRPE + w_2 VPE \tag{4b}$$
$$m2c\text{: } utility_{accept} = w_1 sRPE + w_2 VPE + w_3 APE \tag{4c}$$

Results show that the best-fitting model incorporates empirical reward and valence PEs (m1b; Supplementary Figure 8 and Supplementary Table 15). Comparing m1a BICs (RPE) to m1b BICs (RPE + VPE) shows that model fits are improved by adding valence PEs, $t(72) = -3.55$, $p < .001$. As in Experiment 1, we find that models with empirical RPEs outperform their counterparts with sRPEs (all $p < .001$). We also find that m1b provides a better fit than m1c, $t(72) = -5.98$, $p < .001$, indicating that at the individual-level arousal PEs may not be consistently predictive of choice. Comparing individual fits for m1b and m1c reveals that 20.5% of individuals were best fit by model m1c, suggesting variability in use of arousal PEs.

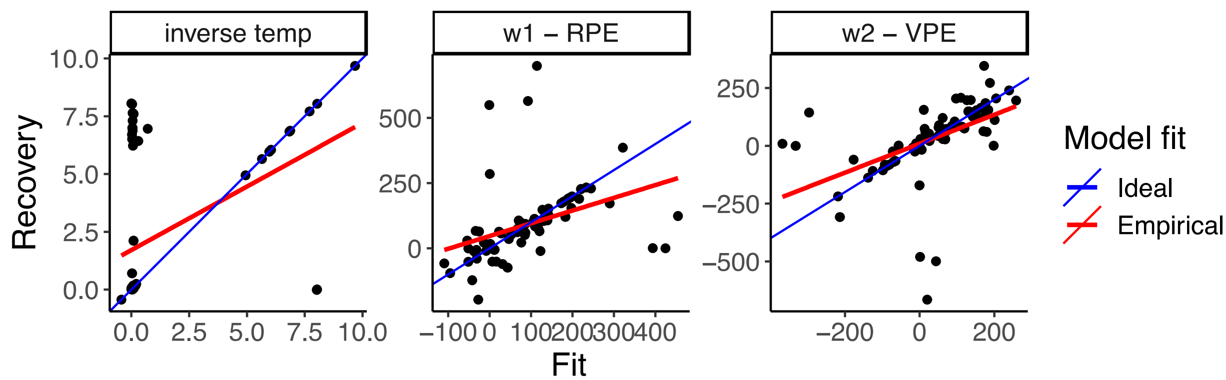


Supplementary Figure 8. Model comparisons for different PE models for Experiment 3 (Justice Game). *The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Individual points represent subject-level BIC values, while the red bars represent the group average. Results show that the additive model (m1b) as the winning model.*

Model	Parameters	Model BIC (± SEM)	BIC – BIC_{model_1b}
1a	2	*23.8 (± 0.64)*	2.23
1b	3	*21.6 (± 0.83)*	*N/A*
1c	4	*23.0 (± 0.87)*	*1.45*
2a	3	*26.5 (± 0.62)*	*4.97*
2b	4	*24.3 (± 0.83)*	*2.79*
2c	5	*25.7 (± 0.88)*	*4.14*

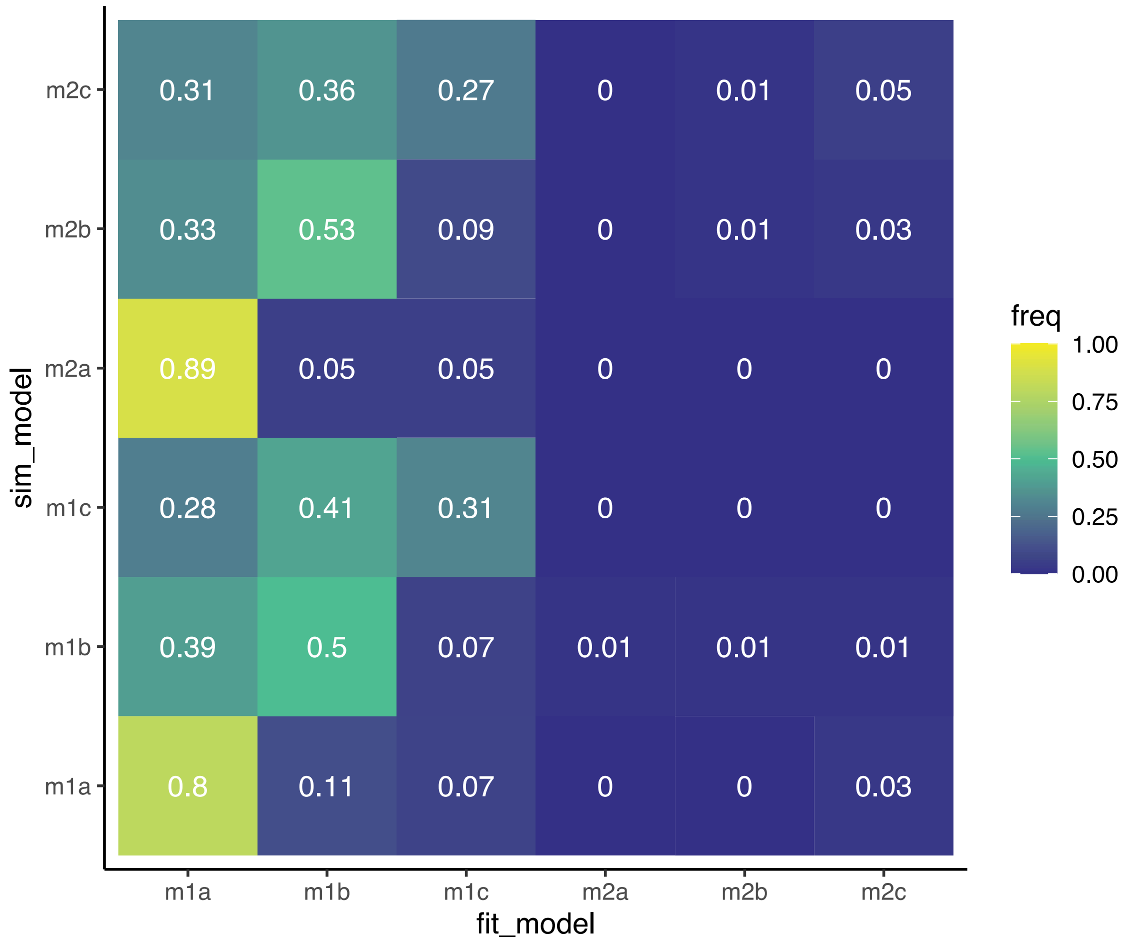
Supplementary Table 15. Model comparisons for different PE models for Experiment 3 (Justice Game). *The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Model BIC shows the average BIC for the model alongside the standard error of the mean.*

Parameter Recovery. We then performed parameter recovery on the winning model (m1b). When comparing the recovered parameters to their true values, we found some evidence for successful parameter recovery using our model (Supplementary Figure 9). The Spearman correlations between the fitted and recovered parameters all show that information is being successfully captured by the model (inverse temp r = .45; w1/rPE r = .70; w2/vPE r = .79). The strength of the correlations between the fitted and recovered parameters shows we can recover this model's parameters despite the relatively low number of trials (18) per participant.



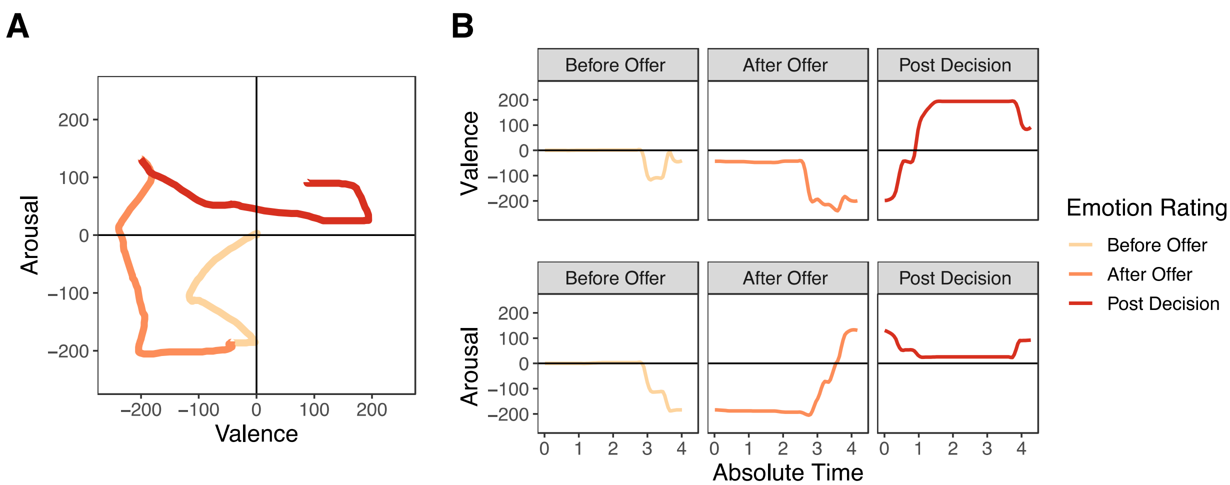
Supplementary Figure 9. Parameter recovery of m1b for Experiment 3 (Justice Game). *Black dots represent individual subjects while the blue and red lines represent the ideal and empirical model fits. Each panel shows the recovery for one of the four parameters included in the model.*

Model Identifiability. Finally, we also examined model identifiability for each of the six models by simulating data using subject-level PEs and the best fitting parameters for each model for Experiment 3. The simulated data was then fit to each model separately, and the best fitting model was assessed with BIC. The results are summarized below in the confusion matrix (Supplementary Figure 10). Results show that m1a (RPE) is highly recoverable, though there is some confusion with m1b (RPE + VPE) and m1c (RPE, VPE, and APE), which provides strong evidence for our hypothesis. Even when the data was generated by m1b or m1c (which incorporate emotion PEs), there is a bias for the model-fitting process to favor m1a (which only incorporates reward PEs). Despite this bias, our model comparison still reveals that m1b (incorporating all reward and valence PEs) provides the best fit to participants' data, revealing that our effect is robust enough to overcome the bias in the model-fitting process.



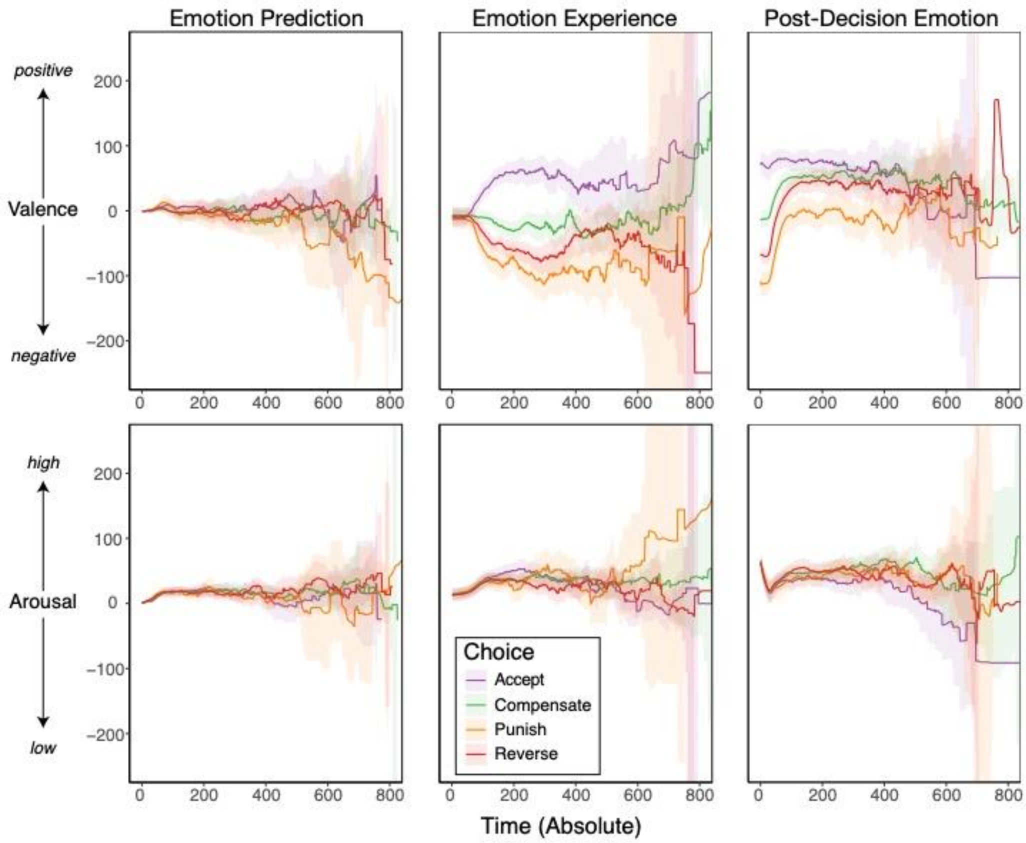
Supplementary Figure 10. Model Identifiability for Experiment 3 (Justice Game). *Confusion matrix for model recovery in the JG. Choices were simulated according to each model and then re-fit according to all six models. Darker colors indicate a lower probability that the fit model is the best model for the simulated data. Lighter colors represent cases with high probabilities.*

Temporal Dynamics Data Preprocessing and Illustration. On each trial, participants made three emotion ratings: before the offer, after the offer, and after making a decision. The task was designed such that the final emotion rating on a specific part of the task (e.g., emotion prediction) would be the starting point for the emotion rating on the next part of the task (e.g., emotion experience). An illustrative participant's data is shown below, which is shown holistically on the 2D dARM (Supplementary Figure 11A), and also for each emotion dimension separately (Supplementary Figure 11B). Adapting instructions from similar mouse-tracking paradigms (Freeman et al., 2011), participants were instructed to respond as quickly and accurately as possible. This helped ensure mouse tracking trajectories were smooth.



Supplementary Figure 11. A) Example participant's emotion trajectory using dARM. *An example participant's data illustrating mouse-tracking trajectory for one Compensate/Reverse decision. This representation shows how the final emotion rating on one measure (e.g., before an offer) becomes the starting point for the subsequent emotion rating (e.g., after an offer).* ***B) Example participant's emotion trajectory separated into distinct dimensions****. The same trial data broken into the valence and arousal dimensions over time. On this trial, the participant chose to Reverse the highly unfair offer.*

Participants' emotion ratings can be represented on two different timescales, an absolute timescale and a normalized timescale. The absolute timescale binned participant's ratings to the nearest 10ms window, and represents individual trajectories through the emotion space on the dARM. However, because trials were self-paced, some trials took longer than others. As a consequence, the absolute timescale results in fewer datapoints at longer time windows (Supplementary Figure 12). Therefore, for the main analysis participants' data were normalized to 101 time bins, where 1 represents the start of a trial and 101 represents when the participant made their rating.



Supplementary Figure 12. Absolute timeseries for emotional experiences during choice. *Participants used the dARM to continuously report their emotional experiences at every stage of the Justice Game. Participants' data are plotted in absolute time, where each unit represents 10ms.* ***Emotion prediction****. The average valence and arousal are plotted over time.* ***Emotion Experience****. The average valence and arousal experience are plotted over time.* ***Post-decision emotion.*** *After making their choice, participants' average valence and arousal experience are plotted for all options. All shaded areas represent 95% CIs.*

Permutation Testing. We examined how real-time fluctuations in emotion bias social behaviors by conducting one-way ANOVA's comparing Compensate choices (group 1) to Reverse choices (group 2) across all trials and at each normalized time bin. Separate tests were performed for the valence and arousal dimensions. As this produced a large number of comparisons across the timecourse (therefore inflating the false positive rate), we adapted nonparametric statistical methods for analyzing EEG timecourse data, which suffer from similar problems (Maris & Oostenveld, 2007). After identifying clusters of significant differences between the groups, we used 1,000 Monte Carlo runs to randomly permute choice labels, run analyses on the resulting permutation data, and store the maximum cluster for each effect on each run. This generates a null-hypothesis distribution which we can compare to our empirical cluster. To summarize, this method allowed to test when significant clusters emerged in the valence and arousal ratings of individuals who chose to *Compensate* or *Reverse*, regardless of the other option presented on a given trial (Figure 3 in the manuscript), while avoiding false positives from multiple comparisons. We can also examine specific option pairs such as when Compensate versus Reverse are presented on the same trial (see manuscript).

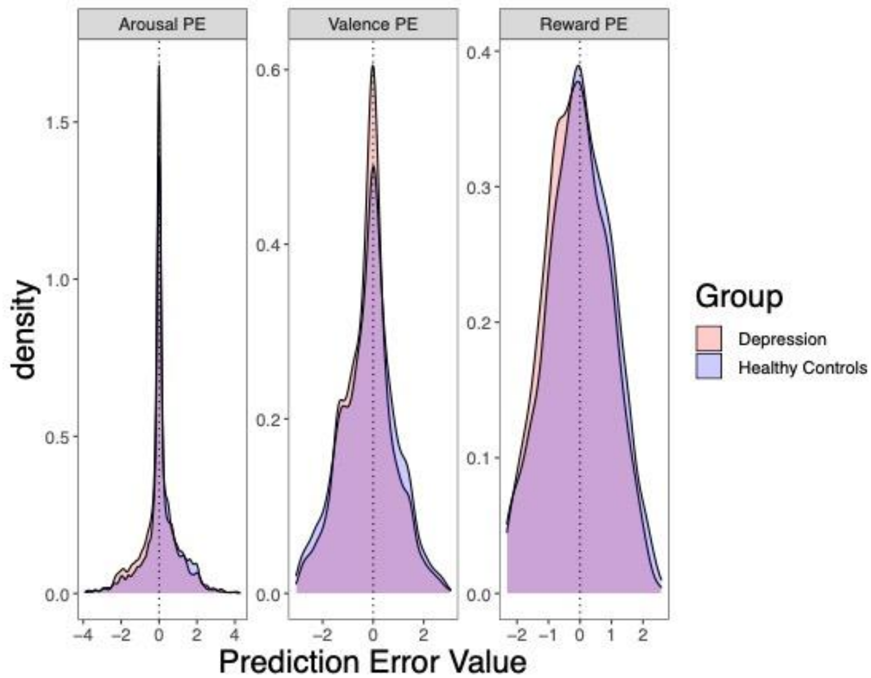
Using emotion trajectories to predict choice. To determine how quickly we could predict a given choice given the trajectory of emotional valence ratings as they unfold over time, we calculated the earliest cluster for each unique trial type (combination of choice pairings and unfairness level) for the emotion experience rating. The table below details the first significant cluster for any given choice pair in both normalized and absolute time units (converted from normalized time unit and average reaction time; Supplementary Table 16). For 15 out of 18 unique trials the emotion trajectories of each choice significantly diverge at some point and three of these diverge before 1s on the absolute time scale.

	Accept-Punish	Accept-Compensate	Reverse-Accept	Punish-Compensate	Reverse-Punish	Compensate-Reverse
Low unfair	75th (3.28s)	NA	54th (1.87s)	23th (0.84s)	30th (2.80s)	30th (2.61s)
Medium unfair	45th (1.57s)	21st (0.77s)	60th (2.27s)	NA	59th (2.20s)	59th (2.06s)
High unfair	59th (2.14s)	28th (0.95 s)	57th (1.95s)	NA	44th (1.46s)	44th (1.48s)

Supplementary Table 16. Using emotional trajectories to predict choice. *Earliest significant clusters were identified using the normalized time trajectories and the first bin where valence ratings significantly differentiated which choice was subsequently taken, given the unfairness level. These bins were then converted to absolute time using the longest average emotion measure reaction time for each pair (e.g., if Compensate took longer than Accept on average, the absolute time uses Compensate).*

2.4 Experiment 4 – Mood disorder and Ultimatum Game Results

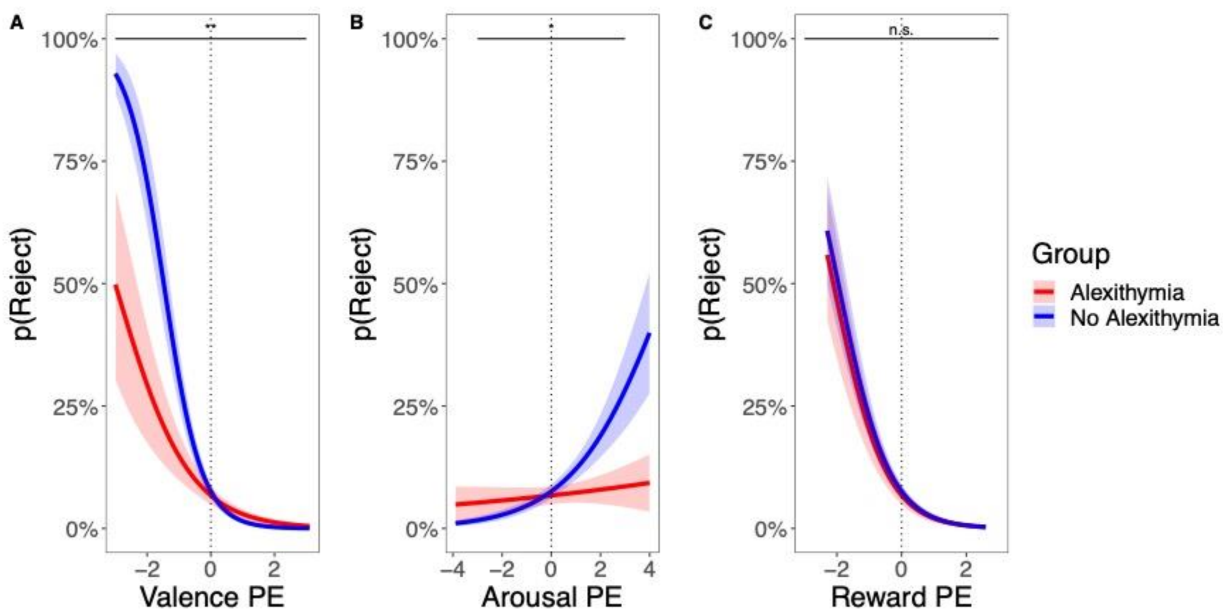
Prediction errors variability across groups. We examined the distribution of prediction errors across groups to determine whether differences in prediction errors underlie selective impairment of emotion prediction error use for individuals at risk for depression. Examining the distribution in valence, arousal, and reward prediction errors for both groups shows that the distributions highly overlap (Supplementary Figure 13). Furthermore, the variability of prediction errors is extremely similar between the healthy controls (Arousal PE SD = 0.99, Valence PE SD = 1.02, Reward PE SD = 1.01) and those that present with depressive symptoms (Arousal PE SD = 1.00, Valence PE SD = 0.93, Reward PE SD = 0.98).



Supplementary Figure 13. Distributions of emotion prediction errors indexed by group. *Density plots are shown separately for arousal, valence, and reward PEs by group (red = depression risk; blue = healthy controls). Purple represents the overlap between the distributions. Prediction errors have been scaled but not mean-centered, as the 0 point on each scale refers to the meaningful instance where expectations match experience.*

Alexithymia predicts selective impairment for emotion, but not reward, prediction errors.

Alexithymia is a sub-clinical disorder characterized by an impairment in identifying and describing emotions in both the self and other (Bagby et al., 1994). Our sample included 90 individuals who met the threshold for alexithymia, 54 individuals possibly meeting threshold for alexithymia, and 207 individuals who did not meet the threshold. For the purposes of this analysis, we only compared individuals who met threshold and those who did not. Using a logistic mixed-effects model, we tested whether the effect of emotion and reward prediction errors on choice change as a function of alexithymia. Results show that individuals who meet this threshold show significant impairments in valence and arousal prediction errors but not in reward prediction error use (Supplementary Table 17; Supplementary Figure 14). Although depression and alexithymia scores were unsurprisingly correlated (r(349) = 0.61, p < .001), this replicates the depression results reported in the manuscript.



Supplementary Figure 14. Alexithymia impairs emotion but not reward prediction errors. *The relationship between alexithymia, choice and all three prediction errors is shown. Plot lines reflect parameter fits based on trial-wise mixed-effects regressions found in Supplementary Table 17. Shaded error bars reflect ±1 SEM.*

Experiment 4: Alxiethymia impairs emotion but not reward prediction errors

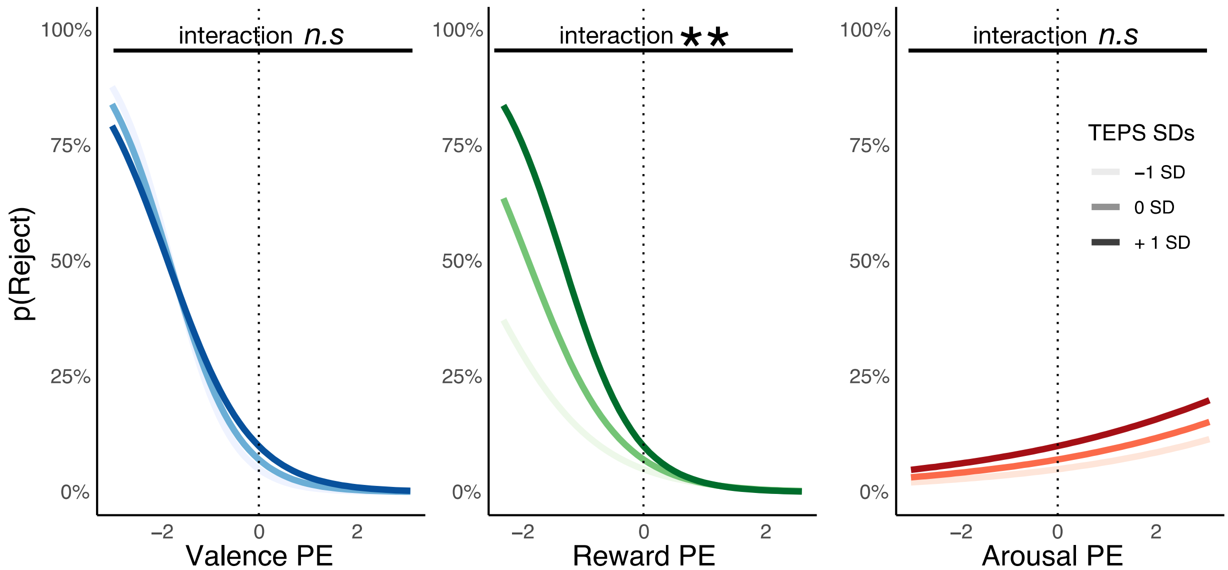
	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.50	0.20	-12.22	**<0.001**
Valence PE	-1.84	0.22	-8.54	**<0.001**
Alexithymia	-0.13	0.33	-0.38	0.706
Arousal PE	0.52	0.13	3.95	**<0.001**
Reward PE	-1.29	0.19	-6.71	**<0.001**
Valence PE:Alexithymia	0.89	0.32	2.79	**0.005**
Arousal PE:Alexithymia	-0.43	0.21	-2.09	**0.036**
Reward PE:Alexithymia	0.03	0.28	0.12	0.904
N_{sub}	297			
Observations	5940			

Supplementary Table 17. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. Alexithymia is a binary variable with no alexithymia (0) and alexithymia (1). The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs.*

Impairments in experiences of pleasure uniquely predict reward—but not emotion—prediction error use.

The temporal experience of pleasure scale (TEPS) is a pleasure scale indexing elements of reward processing. Using the average TEPS score (averaging all items together), we tested whether individuals who scored high on the impairment of pleasure scale showed selective impairments in their use of reward prediction errors. Results showed that individuals with high TEPS scores showed selective reliance of reward PEs, but not emotion PEs (Supplementary

Table 18; Supplementary Figure 15). The TEPS does have two subscales relating to impairments in anticipatory and consummatory pleasure and when we conduct the analysis using those scores we see that the impairments in reward prediction errors are primarily driven by the consummatory pleasure subscale.



***Supplementary Figure 15. Temporal Experience of Pleasure Scale (TEPS) shows a dissociation between reward and emotion prediction errors**. The relationship between TEPS scores, choice and all three prediction errors predicting is shown. For visualization purposes, the TEPS was split into three categorical values (-1 SD, 0, SD, + 1 SD), but the significance tests reflect the full continuous interaction.*

Experiment 4: TEPS shows impairments in reward, but not emotion, prediction errors

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.58	0.17	-15.47	**<0.001**
Valence PE	-1.53	0.16	-9.62	**<0.001**
TEPS	0.38	0.14	2.62	**0.009**
Arousal PE	0.27	0.11	2.61	**0.009**
Reward PE	-1.38	0.14	-9.84	**<0.001**
Valence PE:TEPS	0.25	0.14	1.83	0.067
Arousal PE:TEPS	-0.02	0.09	-0.18	0.858
Reward PE:TEPS	-0.31	0.12	-2.63	**0.008**
N_{sub}	351			
Observations	7020			

Supplementary Table 18. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. TEPS is a continuous score averaging all the items. The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs.*

Other reward-related questionnaires and their relationship with PE use.

We tested whether the other reward-related questionnaires were also linked with impaired reward PE use, including the BIS/BAS (Carver & White, 1994), apathy evaluation scale (Marin et al., 1991), and the Snaith-Hamilton Pleasure Scale (Snaith et al., 1995). Separate regression analyses reveal a marginal interaction between the BIS subscale and valence PEs (B = 0.27 (0.15), z = 1.78, p = 0.08), no interactions between the AES and any PEs (ps > .31), and no interactions between the SHAPS and any PEs (ps > .30).

Use of valence and arousal prediction errors is impaired in depression. We analyze all participants together in the manuscript (Table 5) and show separate analyses for healthy controls (Supplementary Table 19) and individuals at risk for depression (Supplementary Table 20).

Experiment 4: Healthy Controls use of PEs

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-3.00	0.27	-11.17	**<0.001**
Valence PE	-2.08	0.29	-7.15	**<0.001**
Arousal PE	0.57	0.18	3.21	**0.001**
Reward PE	-1.41	0.23	-6.03	**<0.001**
N_{sub}	205			
Observations	4100			

Supplementary Table 19. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs. Data only includes individuals not at risk for depression.*

Experiment 4: Individuals at risk of depression use of PEs

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-2.21	0.21	-10.34	**<0.001**
Valence PE	-0.94	0.16	-5.81	**<0.001**
Arousal PE	0.01	0.13	0.09	0.930
Reward PE	-1.43	0.18	-7.79	**<0.001**
N_{sub}	146			
Observations	2920			

Supplementary Table 20. *Reward PEs are calculated as the difference between the experienced and predicted reward. Valence and Arousal PEs are calculated as the difference between the experienced and predicted emotion on each dimension. All variables were scaled but not mean-centered, as the 0 point refers to the meaningful instance where expectations match experience. The model includes subject-specific random intercepts and random slopes for Reward, Valence, and Arousal PEs. Data only includes individuals at risk for depression.*

Experienced emotions in the depressed group. Using a logistic mixed-effects model, we tested whether the effect of experienced reward and emotion changes as a function of being at risk for depression. Results show that while both groups report similar experiences of reward, individuals at risk for depression experienced less arousal compared to healthy controls (p = .074; Supplementary Table 21). This was mirrored more strongly for the experience of emotional valence.

Experiment 4: Clinical groups use of experience

	Estimates			
Predictors	*Log-Odds*	*SE*	*Z*	*p*
Intercept	-5.05	0.45	-11.24	**<0.001**
Valence Experience	-2.55	0.33	-7.71	**<0.001**
Depression	1.64	0.54	3.06	**0.002**
Arousal Experience	0.22	0.22	0.98	0.328
Reward Experience	-3.73	0.39	-9.66	**<0.001**
Valence Experience:Depression	1.29	0.40	3.20	**0.001**
Arousal Experience:Depression	-0.51	0.28	-1.79	0.074
Reward Experience:Depression	0.66	0.44	1.50	0.135
N_{sub}	351			
Observations	7020			

Supplementary Table 21. *Experienced Reward is defined as the amount of money offered by Player A and was normalized (scaled and mean-centered). Experienced Valence and Arousal are the experienced emotions on each dimension when receiving the offer and were scaled but not mean-centered (as the 0 point on each scale represents a neutral feeling). Depression is a binary variable with healthy controls (0) and individuals not at risk of depression (1). The model includes subject-specific random intercepts and slopes for Experienced Reward, Experienced Valence, and Experienced Arousal.*

Expectation and Prediction Error Model Comparisons. To better understand how expectations and prediction errors contributed to choice, we compared models which contained expectation and PE terms for reward only (m4a), reward and valence (m4b), and reward, valence, and arousal (m4c) for the healthy controls in Experiment 4. Likelihood Ratio Tests reveal that a model which only includes reward expectations and reward PEs (m4a) fit the data worse than a model which added valence expectations and valence PEs (m4b; $\chi^2(2) = 100.58,\ p < .001$). While adding arousal expectations and arousal PEs (m4c) did not improve the model fit, m4b was also not significantly better than m4c ($\chi^2(2) = 3.58,\ p = .17$), as they had comparable BICs (m4b: 1708.0, m4c: 1721.1).

Additive and Interactive Model Comparisons. We compared models which contained all three prediction errors additively to an interactive model for healthy controls. Likelihood Ratio Tests reveal that a model which includes interactions between prediction errors fits better than a model with no interactions ($\chi^2(4) = 34.02,\ p < .001$).

Controlling for Expectations in the Interactive Model. We used nested mixed-effects regressions to test whether the interactions between prediction errors still hold when controlling for reward, valence, and arousal expectations for the healthy controls. We compared two models: one which included all PEs and their interactions (the “interactive” model), and one which adds expectations (reward, valence, and arousal) on top of the interactive model. Likelihood ratio tests reveals that adding expectations improves the fit of the data $\chi^2(3) = 928.4,\ p < .001$. In this model, both valence and reward PEs remain significant after accounting for expectations, although arousal PEs is only marginally significant (Supplementary Table 22).

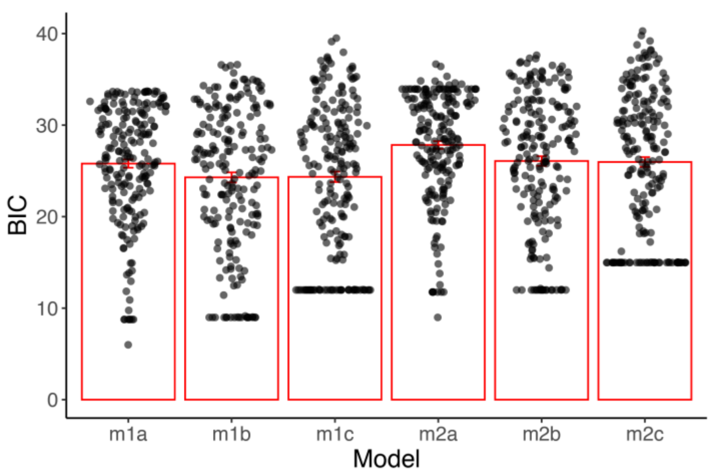
Experiment 4: Controlling for Expectations in the Interactive Model

Predictors	Estimates			
	Log-Odds	*std. Error*	*Statistic*	*p*
Intercept	2.01	0.65	3.11	**0.002**
Valence Expectations	-1.81	0.26	-7.10	**<0.001**
Arousal Expectations	0.36	0.19	1.88	0.060
Reward Expectations	-8.40	0.74	-11.37	**<0.001**
Valence PE	-2.74	0.30	-9.02	**<0.001**
Arousal PE	0.14	0.21	0.68	0.498
Reward PE	-5.57	0.48	-11.67	**<0.001**
Valence PE: Arousal PE	-0.03	0.20	-0.16	0.873
Valence PE: Reward PE	-0.43	0.18	-2.41	**0.016**
Arousal PE: Reward PE	-0.21	0.20	-1.06	0.287
Valence PE: Arousal PE: Reward PE	-0.15	0.13	-1.20	0.231
N_{sub}	205			
Observations	4100			

Supplementary Table 22. *All variables were scaled but not mean-centered, as the 0 points are meaningful. The model includes subject-specific random intercepts and random slopes for Reward PEs.*

Computational Modeling Results. We repeated the formal modeling techniques described in Experiment 1 for the healthy controls in Experiment 4. Models labeled with "*m1*" use empirical reward prediction errors (defined as monetary reward, minus monetary reward prediction), while models labeled with "*m2*" use subjective reward prediction errors (defined above).

Results show that the subjective reward PE model (*m2a*)—which incorporates the non-linear valuation of reward into the prediction error itself—provides a worse fit than the model that uses empirical reward PEs (*m1a*; Supplementary Figure 16; Supplementary Table 23; *Model 2a vs 1a: t(204) = 26.02, p < .001*). Although both model fits are improved by adding valence PEs to models with only reward PEs (*Model 1b vs 1a: t(203) = -4.230, p < .001; Model 2b vs 2a: t(204) = -4.72, p < .001)*), we again find that the model using empirical RPEs (*m1b*) outperforms the model using subjective RPEs (*m2b*; *t(204) = -10.55, p < .001*). Finally, we find that adding arousal in addition to reward and valence PEs does not provide a better fit and models omitting arousal PEs fit best (*Model 1c vs 1b: t(204) = 0.21, p = .84; Model 2c vs 2b: t(204) = -0.34, p = .73*). However, these results show that the two models produce similar BICs. Comparing individual fits for m1b and m1c revealed that 31.7% of participants were better fit by m1c, highlighting the potential variability in use of arousal PEs. Finally, again a model with empirical RPEs (*m1c*) out performs the corresponding model using subjective RPEs (*m2c; t(204) = -8.91, p < .001).*

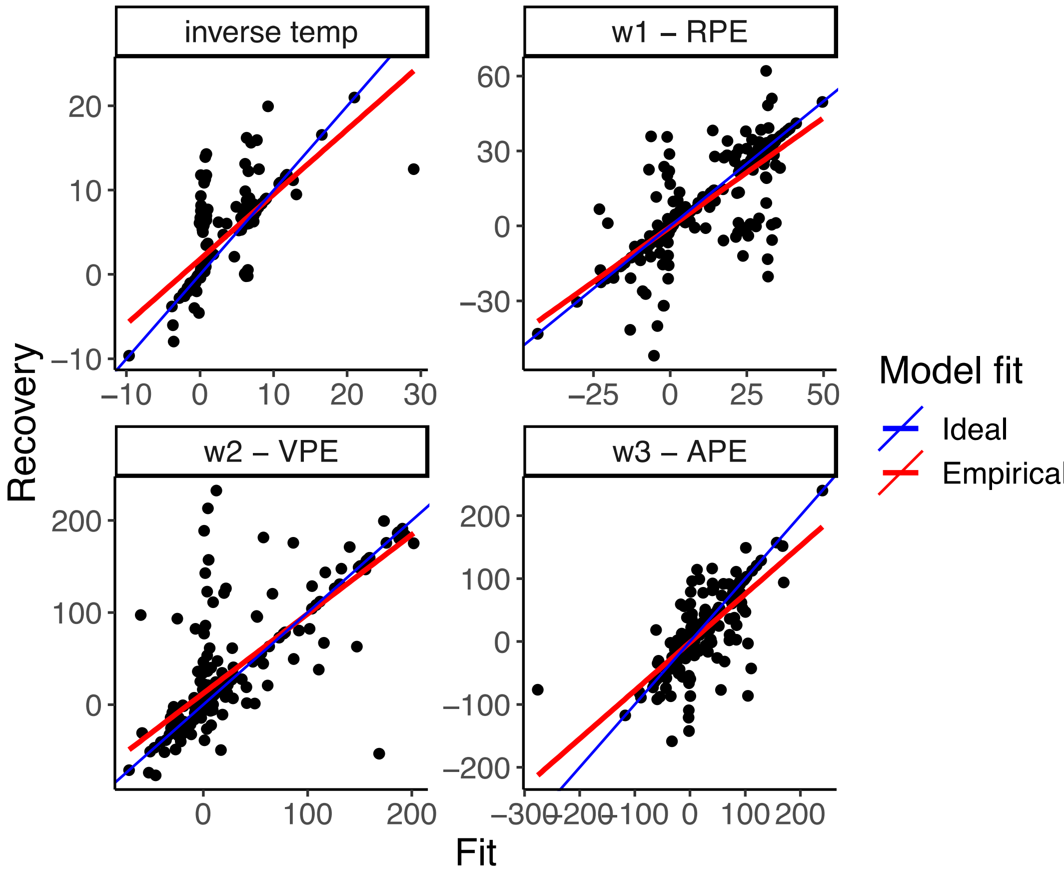


Supplementary Figure 16. Model comparisons for different PE models. *The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Individual points represent subject-level BIC values, while the red bars represent the group average.*

Model	Parameters	Model BIC (± SEM)	BIC – BIC_{model_1b}
1a	2	*25.80 (± 0.45)*	1.51
1b	3	*24.29 (± 0.55)*	*N/A*
1c	4	*24.35 (± 0.57)*	*0.06*
2a	3	*27.83 (± 0.42)*	*3.54*
2b	4	*26.08 (± 0.52)*	*1.79*
2c	5	*25.97 (± 0.54)*	*1.68*

Supplementary Table 23. Model comparisons for different PE models. *The models using empirical reward PEs (i.e., those originally reported in our manuscript) are labeled as "m1", while the models that use subjective reward PEs (i.e., nonlinear transformations of reward into subjective utility) are labeled as "m2". The a, b, and c variants of the models indicate the stepwise inclusion of various PEs into the model. Models with "a" only have reward PEs, models with "b" have reward PEs and valence PEs, and models with "c" have all three PEs (reward, valence, and arousal). Model BIC shows the average BIC for the model alongside the standard error of the mean.*

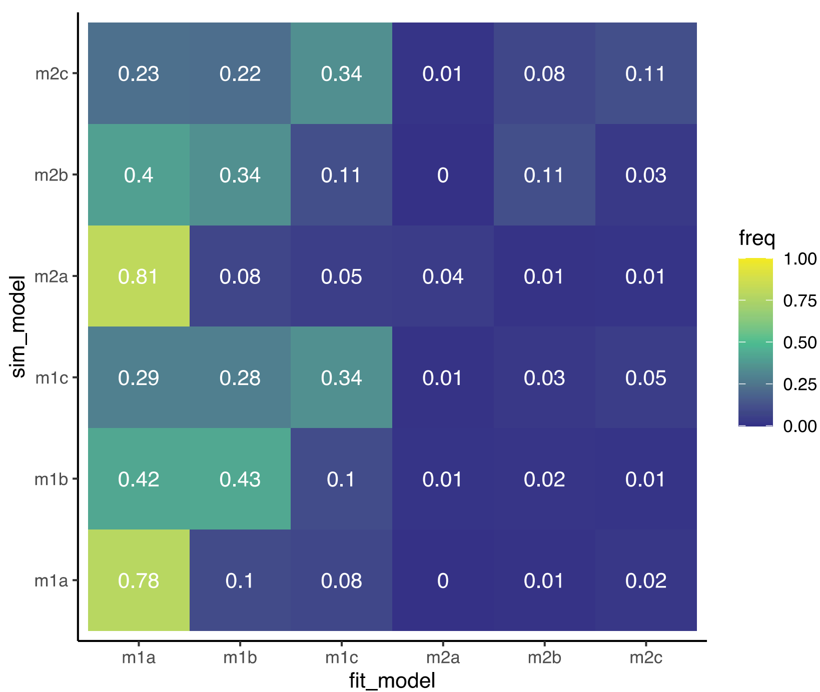
Parameter Recovery. We performed parameter recovery on model (*m1c*, which had similar BICs to m1b) by first identifying the best-fitting parameters for that model, which include weights on each prediction error (w1, w2, and w3 for reward, valence, and arousal PEs, respectively), as well as an inverse temperature term (beta) in the softmax choice rule. Once we identified the best-fitting parameters, we simulated data with these known parameter values, and re-fit the model on the simulated data (Nassar, Waltz, Albrecht, Gold, & Frank, 2021). When comparing the recovered parameters to their true values, we found evidence for successful parameter recovery using our model (Supplementary Figure 17). The Spearman correlations between the fitted and recovered parameters all show that information is being successfully captured by the model (inverse temp r = .75; w1/rPE r = .77; w2/vPE r = .78; w3/aPE r = .75). The strength of these correlations shows we can recover this model's parameters despite the relatively low number of trials (20) per participant.



Supplementary Figure 17. Parameter recovery of m1c. Black dots represent individual subjects while the blue and red lines represent the ideal and empirical model fits. Each panel shows the recovery for one of the four parameters included in the model.

Model Identifiability. We also examined model identifiability for each of the six models by simulating data using subject-level PEs and the best fitting parameters for each model. The simulated data was then fit to each model separately, and the best fitting model was assessed with BIC. The results are summarized below in the confusion matrix (Supplementary Figure 18), where the numbers correspond to the percentage of simulations for which the model from the empirical data (x-axis) best fit the simulated model's data (y-axis). Perfect recovery would yield

0 everywhere but along the diagonal (which should have a value of 100)—although in practice, this is rarely the case (e.g., Wilson & Niv, 2012). Results show that model 1a (rPE only) is highly recoverable, though there is some confusion between this model and model 1b (rPE + vPE) and 1c (rPE, vPE, and aPE). We interpret this as providing even stronger evidence for our hypothesis. Even when the data was generated by m1b or m1c (which incorporates emotion prediction errors), there is a bias for the model-fitting process to favor m1a (which only incorporates reward PEs). Despite this bias, our model comparison still reveals that m1c (incorporating all three PEs) provides the best fit to participants' data, implying that our effect is robust enough to overcome the bias in the model-fitting process.



Supplementary Figure 18. Model Identifiability for Experiment 4. *Confusion matrix for model recovery in the UG. Choices were simulated according to each model and then re-fit according to all six models. Darker colors indicate a lower probability that the fit model is the best model for the simulated data while lighter colors represent higher probabilities.*

Chapter 5

Dissociable neural and temporal dynamics of emotion and reward prediction errors

Joseph Heffner & Oriel FeldmanHall
In Preparation

Introduction

The reward prediction error (PE)—the difference between experienced reward and expected reward—is the dominant construct for explaining how people learn to make adaptive decisions (Daw et al., 2011; King-Casas et al., 2005; Pessiglione et al., 2006; Rouhani & Niv, 2021; Schultz et al., 1997; Schultz & Dickinson, 2000). A class of computational accounts of these behaviors, called reinforcement learning (RL) models (Collins & Shenhav, 2022; Sutton & Barto, 2018), offers a normative framework of how people can achieve optimal decision-making by learning to predict events in the future and choose actions to maximize rewards through reward PEs (Niv & Montague, 2009; Silver et al., 2021). This unifying framework for all decision-making has been fruitfully applied to explain learning simple behaviors, such as learning to avoid financial losses (Pessiglione et al., 2006) or navigating novel environments (Sosa & Giocomo, 2021), and even complex social behaviors, such as determining who can be trusted (Lamba et al., 2020) or teaching others novel environments (Ho et al., 2017). Despite the ubiquity of RL models in decision-making, virtually all accounts ignore one of the most consistent factors implicated in learning and choice—emotion.

In a separate, but related, literature, emotions have been widely recognized as an important factor in the decision-making process (Lerner et al., 2015; Phelps, 2005), as studies have shown how stress (FeldmanHall et al., 2015), inducing mood (Stanton et al., 2014) and regulating emotions (Martin & Delgado, 2011) can all alter choices. Indeed, almost all emotion theories characterize some relationship between emotions, thought of as evaluations of external stimuli (e.g., rewards), and determinants of future behavior (e.g., action tendencies; Frijda, 1987; Lowe, 2011). In a first step toward conceptualization emotion's role in a reinforcement learning framework, we recently demonstrated that violations of emotion expectations—emotion prediction errors (PEs)—have an independent and dissociable role from RPEs in predicting social choices such as a willingness to punish a transgressor for unfair behavior (Heffner et al., 2021). Furthermore, we showed a functional dissociation such that individuals at risk of depression showed selective impairments in emotion PEs predicting choices, but not reward PEs. These results demonstrated the central role of emotion PEs for guiding social behavior, but it remains unknown whether emotion and reward PEs have separable effects on learning. As such, two outstanding questions remain: 1) how do emotion PEs contribute to learning about social partners and predicting choices over time, and 2) are emotion and reward PEs dissociable at the neural level?

In the present study, we test both questions using a social learning paradigm and electroencephalography (EEG). We chose EEG because of its excellent temporal resolution (on the order of milliseconds), which would allow us to capture whether reward and emotion PEs are processed at different time scales during learning and decision-making. To test for neural dissociability between reward and emotion PEs, we identified three candidate event-related potentials (ERPs) that are changes in EEG activity in response to feedback or reward processing. The main ERP associated with reward learning is the feedback-related negativity (FRN), an ERP that is sensitive to surprising events (Hauser et al., 2014; Hayden et al., 2011) and is theorized to directly reflect RPE processing (Holroyd & Coles, 2002). The other candidate ERPs, the P3a and P3b, have been associated with many aspects of feedback processing including the magnitude of reward (Yeung & Sanfey, 2004), the valence of reward (Hajcak et al., 2005), and rare, surprising outcomes (Mars et al., 2008). We hypothesized that if a neural dissociation is to be found between emotion and reward PEs that these components are the most likely candidates, but we had no clear predictions about which signals would preferentially index reward or emotion processing.

We tested these predictions using behavioral and EEG data from 35 participants learning about social partners during a repeated Ultimatum Game (UG). The UG is a simple negotiation game where a proposer is given ten dollars to split with the participant, the responder, who decides whether to accept or reject the division. If the division is accepted, then both earn the amount proposed but if rejected, both earn nothing for that round. Unknown to participants, partners' offers were determined by a computer algorithm to create three quantitively distinct partner types who gave bad, neural, or good offers (see Methods). To make this a learning paradigm, participants play five rounds in a row with an individual before being matched with a new partner. Critically, participants should update their expectations based on offer feedback to learn what types of offer their partner generally gives. Our task included three key measurements that allowed us to better understand how rewards and emotions affect feedback processing and trial-by-trial learning (Fig. 1): (1) Participants were required to predict the offer they would receive on each trial and indicate it on a visual analog scale (reward expectation) and (2) Participants indicated how they thought they would feel after receiving the offer (emotion expectation) and how they actually felt after the offer (emotion experience) on a 2D affect grid used in our prior research (Heffner & FeldmanHall, 2022; Heffner et al., 2021). This measure captures core affect, a consciously accessible aspect of subjective feelings partitioning feelings into core dimensions of valence (pleasurableness) and arousal (alertness/intensity). With these measures, we can calculate three empirical prediction errors (reward, valence, and arousal PEs) as participants learn each trial and test how these measures predict learning, choices, and whether they are indexed by discrete ERPs.

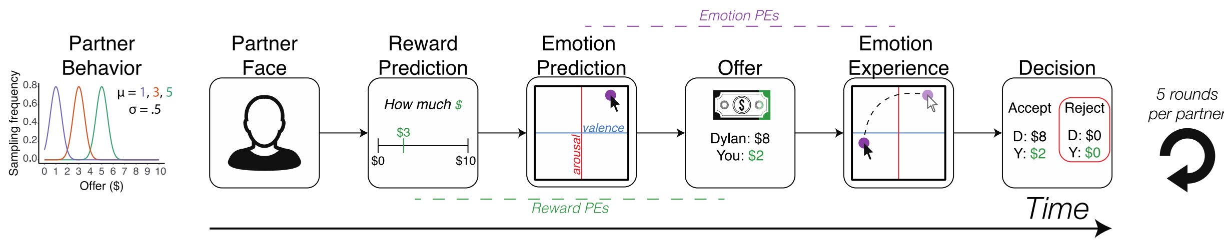


Figure 1. Ultimatum game trial design. *Participants are partnered with another individual and play five rounds of the UG. Unbeknownst to participants, there were three distinct partner types (bad, neutral, and good), whose offers were determined through normal distributions with different means. On each trial, participants made a prediction about how much money they expected to be offered, as well as a prediction about the emotions they expected to experience. After receiving an offer, participants reported the current emotional experience and then decided to either accept or reject the offer.*

Behavioral results

Rapid learning indexed by both reward and emotion prediction errors

How did emotion and reward expectations change as participants learned about their social partners? Results reveal that both reward and emotion PEs predict trial-by-trial learning (Fig. 2). Reward PEs reflect learning about the statistical regularity of their partners' offer (i.e., bad, neutral, or good offers) and one-sample t-tests showed that participants fully learned to predict offers from the bad partner by round five, as reward PEs were not significantly different from 0 ($t(34) = -1.82$, $p = .08$), but there were persistent, optimistic errors as they expected better offers from good ($t(34) = 2.33$, $p = .03$) and neutral ($t(34) = 3.47$, $p = .001$) partners even on the final round. Emotion PEs reflect deviations from how participants will feel about future offers on both the valence and arousal dimensions. The fact that these emotion PEs are no longer predictive by round five for the good (valence: $t(34) = 0.75$, $p = 0.452$; arousal: $t(34) = 1.93$, $p = 0.062$), neutral (valence: $t(34) = 1.77$, $p = 0.086$; arousal: $t(34) = 1.55$, $p = 0.131$), and bad partners (valence: $t(34) = 0.34$ $p = 0.739$; arousal: $t(34) = 3.1$, $p = .004$), suggests that emotion PEs are able to rapidly index learning.

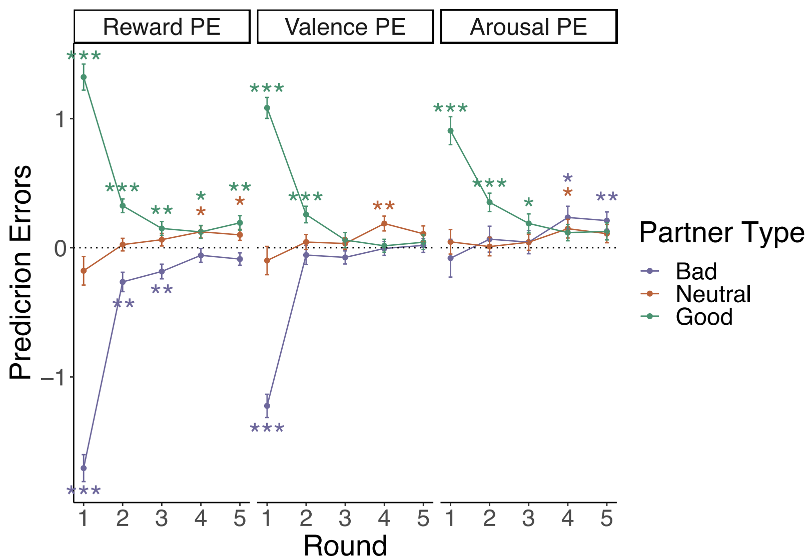


Figure 2. Learning in the ultimatum game. *Prediction errors (PEs) over rounds in the UG. PEs are calculated as the difference between experiences and predictions reported by participants on each round. PEs are averaged within and then across participants for each round and partner type. All error bars reflect ± 1 S.E. ***P<.001, **P<.01, *P<.05.*

Learning about social partners changes how an offer is evaluated

Participants quickly learn about the types of offers their partner will give, but what impact does this have on their choices? Using generalized mixed effects linear regression, we found that participants rejected bad offers (59%) more often than neutral offers (32%; $\beta = 1.91 \pm 0.22, z = 8.76, P < .001$), and in turn rejected neutral offers more often than good offers (9%; $\beta = 2.61 \pm 0.40, z = 6.54, P < .001$; Fig. 3A). The probability of rejecting offers from bad ($\beta = -0.18 \pm 0.04, z = -4.66, P < .001$), and neutral ($\beta = -0.09 \pm 0.04, z = -2.08, P = 0.04$) partners decreased over rounds, suggesting that participants might be changing their decision policy as rejections near the end of the game have no strategic power to influence future offers. However, rejection of offers from good partners actually increased over rounds ($\beta = 0.22 \pm 0.06, z = 3.62, P < .001$; Fig. 3B); one possible explanation is that participants' evaluation of offers changed based on learning about the type of partner they were interacting with. To investigate this, we examined how similar offers were evaluated when coming from different partner types. Figure 3C shows the distribution of bad, neutral, and good partners' offers which overlap between partner types. Generalized linear mixed-effects models showed main effects for partner type, indicating that the same offer (e.g., \$2.15) is rejected less if it came from a bad partner than if it came from a neutral partner ($\beta = -1.88 \pm 0.64, z = -2.93, P = .003$). A similar effect existed between neutral and good partners, such that the same offer (e.g., \$3.50) is rejected less from neutral partners compared to good ($\beta = -3.83 \pm 1.42, z = -2.71, P = .007$). In other words, participants learned to expect what type of offer they were receiving depending on the partner type (bad: \$1, neutral: \$3, good: \$5), and these expectations changed how they evaluated and punished offers.

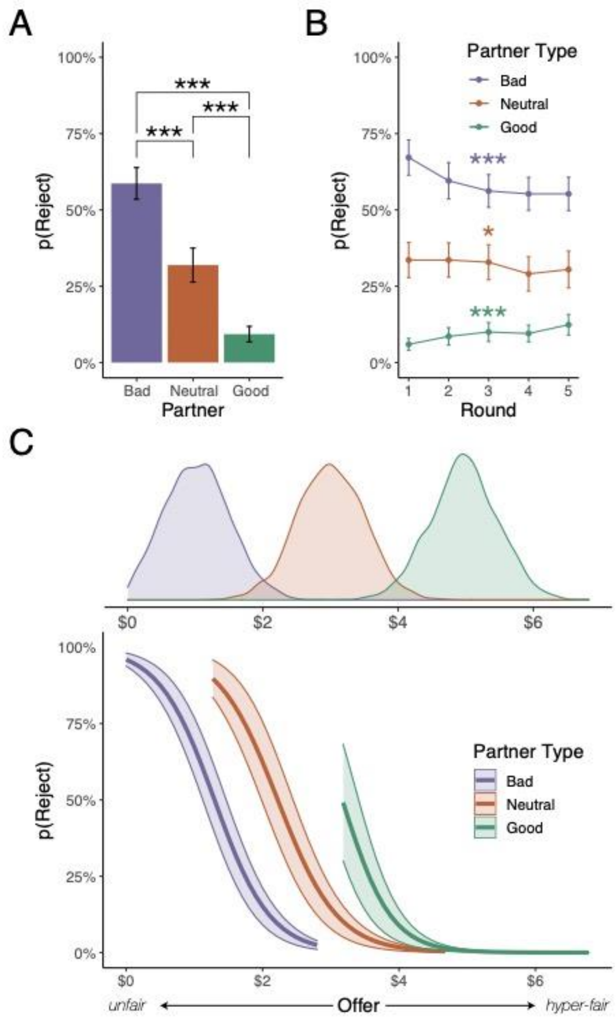


Figure 3. Choice behavior in the ultimatum game. *A) Overall choices per partner type. Bars reflect average proportion of rejections per partner type. B) Choices over round per partner type. Dots reflect average proportion of rejections per round per partner type. Analyses reflect regression effects. C) Offer evaluation by partner type. Histograms reflect empirical distribution of offers for partner type across participants. The probability of rejecting the offer is plotted for all three partner types and analyses represent regression effects. All error bars and shaded areas reflect ±1 S.E. ***P<.001, **P<.01, *P<.05.*

Behavioral separability of valence and reward prediction errors

Our next question focused on whether we could replicate our prior findings—that emotion PEs outperform reward PEs in predicting choices (Heffner et al., 2021)—but in a learning context. To distinguish between the contributions of reward and emotion PEs during the UG learning task, we used an additive model and interacted all PEs with round number to predict choices to punish. Main effects from the generalized linear mixed-effects model reveal that valence and reward, but not arousal, PEs contribute to decisions to punish across rounds (Table 1). That is, participants punished at higher rates when experiencing less reward or pleasantness (valence) than expected. Critically, a beta coefficient test showed that valence PEs have a significantly stronger impact on motivating punitive choices than reward PEs during the first round ($z = -2.10$, $P = 0.02$), replicating our prior work (Heffner et al., 2021).

But do these valence and reward PEs predict choices over time? The interactive terms in the model reveal that valence and reward PEs relationship with choice changes in opposite ways across rounds, such that the effect of valence PEs is strongest at round 1 and weakens over time, while reward PEs grow in their predictive value over time (Table 1 and Fig. 4). These results highlight how valence PEs are more impactful early on, when uncertainty about the social partner and their offers is high. Once participants have more accurate expectations about the type of offer to expect, reward PEs become more useful in informing whether to accept or reject.

Table 1. Separable effects of valence and reward prediction errors predicting choices

$$Punish_{\mathrm{i,t}} \sim \beta_0 + \beta_1 Reward\ PE_{\mathrm{i,t}} + \beta_2 Valence\ PE_{\mathrm{i,t}} + \beta_3 Arousal\ PE_{\mathrm{i,t}} + \beta_4 Round_{i,t} + \beta_5 Reward\ PE_{i,t}: Round_{i,t} + \beta_6 Valence\ PE_{i,t}: Round_{i,t} + \beta_7 Arousal\ PE_{i,t}: Round_{i,t} + \varepsilon$$

Variable	Estimate (SE)	z	p
Punish			
Intercept	-1.37 (0.32)	-4.34	<.001***
Reward PE	-0.47 (0.18)	-2.59	.009**
Valence PE	-1.10 (0.17)	-6.64	<.001***
Arousal PE	-0.02 (0.15)	-0.15	.884
Round	0.03 (0.03)	1.25	.212
Reward PE×Round	-0.08 (0.04)	-2.09	.037*
Valence PE×Round	0.09 (0.04)	2.40	.016*
Arousal PE×Round	0.04 (0.03)	1.30	.194

Note. Reward PEs are calculated by taking the difference between the experienced and predicted reward. Valence PEs and Arousal PEs are calculated by taking the difference between the experienced and predicted emotion. All variables were scaled but not mean-centered, as the 0 point on each scale refers to the meaningful instance where participants' expectations matched their experience. The model includes subject-specific random intercepts and slopes for Reward PE, Valence PE, and Arousal PE. The dataset includes 6,296 observations from 35 participants. *** p <.001.

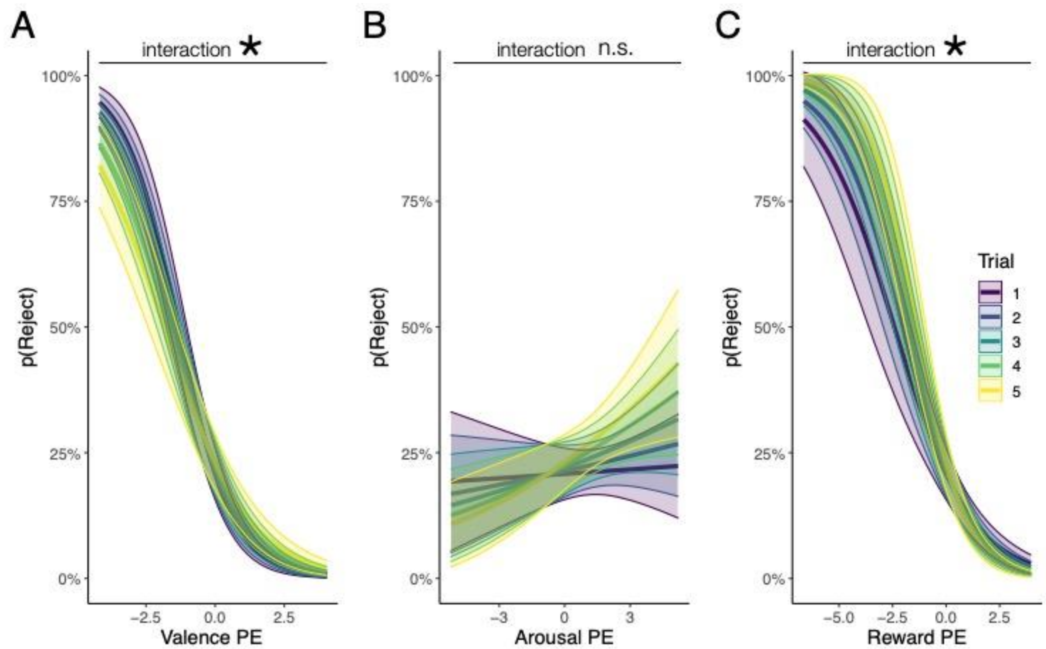


Figure 3. Separable contributions of valence and reward prediction errors for choice. *The probability of rejecting is shown for all three types of PEs (valence, arousal, and reward). The colour of each line indicates the round number (1-5). Negative values reflect negative PEs, indicating less pleasantness (valence), arousal, and money (reward) than expected. Shaded areas reflect ±1 S.E. ***P<.001, **P<.01, *P<.05.*

Neural results

Unsigned reward prediction errors are indexed by the feedback-related negativity

Which PE type does the feedback-related negativity (FRN) track? Previous literature suggests that FRN amplitudes might reflect signed reward PEs (Holroyd & Coles, 2002) or they could resemble reward surprise, which would be captured by absolute reward PEs (Hauser et al., 2014; Hayden et al., 2011). Given this mixed prior literature, we tested both accounts and found that unsigned (absolute) PEs were a better model of FRN response in our paradigm. A linear mixed-effects model predicting trial-by-trial FRN amplitude absolute PEs revealed that only absolute reward PEs robustly predicted FRN amplitudes ($\beta = -0.25 \pm 0.08, t = -3.05, P = 0.004$), while absolute valence PEs were not significant ($\beta = 0.05 \pm 0.08, t = 0.63, P = 0.53$) and arousal PEs (which are already theoretically unsigned) were also non-significant ($\beta = 0.03 \pm 0.05, t = 0.63, P = 0.54$; Fig. 4). In keeping with prior literature's suggestion that the FRN is the neural index of RPEs, we found that FRN amplitude uniquely reflects the degree to which offers deviate from reward expectations and does not reflect violations of emotion expectations.

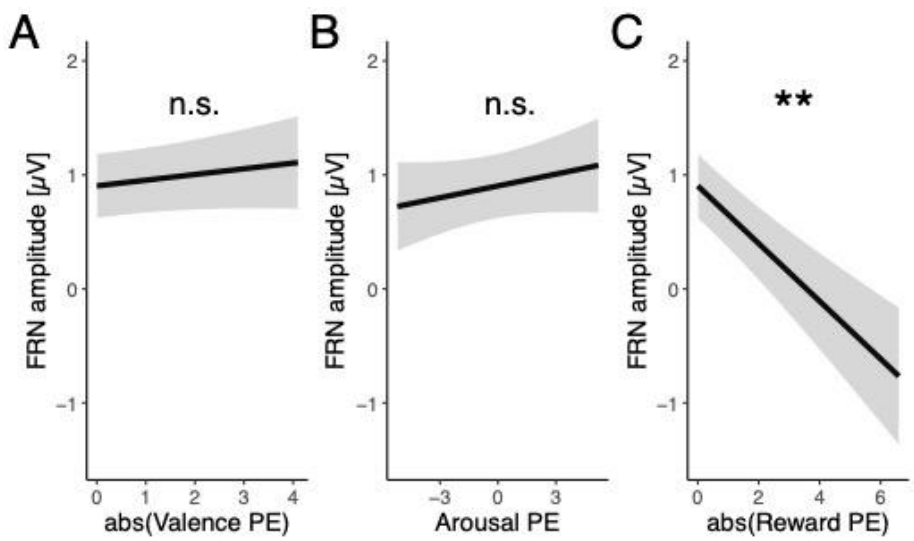


Figure 4. Absolute reward prediction errors (surprise) predict the feedback-related negativity. *Predicted FRN amplitudes are shown for all three types of absolute PEs (absolute valence, arousal, and reward). Negative values reflect less surprise for all three PEs. Analyses reflect regression models and shaded areas reflect ±1 S.E. ***P<.001, **P<.01, *P<.05.*

Arousal prediction errors weakly predict the novelty P3a

We next tested whether any of the absolute PEs predicted trial-by-trial P3a amplitude. The P3a, also known as the novelty P3a (Polich, 2003), has been associated with the engagement of attention to unsigned surprising and motivationally relevant stimuli and functionally is thought to reflect mobilization for action (Nieuwenhuis et al., 2011). Using the same analysis described above, results showed a marginal significant effect for arousal PEs ($\beta = 0.15 \pm 0.07, t = 1.99, P = 0.06$) while absolute valence ($\beta = 0.07 \pm 0.12, t = .62, P = 0.54$) and absolute reward PEs ($\beta = 0.01 \pm 0.11, t = .11, P = 0.91$) were not significant (Figure 5A-C). While this provides some evidence that arousal PEs are indexed through P3a amplitudes, the P3a is likely related to the offer itself, since this is the novel information participants are considering. Because the P3a is a non-valenced component—i.e., it is sensitive to positive and negative surprise—we included a quadratic component in the model to index offer extremity. We found that P3a amplitudes were robustly tracked by the quadratic effect of offer magnitude ($\beta_{quad} = 28.41 \pm 6.41, t = 4.43, P < .001$; Figure 5D), which likely explains the P3a relationship to arousal PEs, as arousal PEs alone correlate with offer extremity ($r_{rm} = .16, P < .001$) while valence PEs ($r_{rm} = -0.01, P = .40$) and reward PEs ($r_{rm} = -0.02, P = .16$) do not. Together, these results demonstrate that arousal PEs, likely driven by offer extremity for both low and high offers, modulate P3a amplitude.

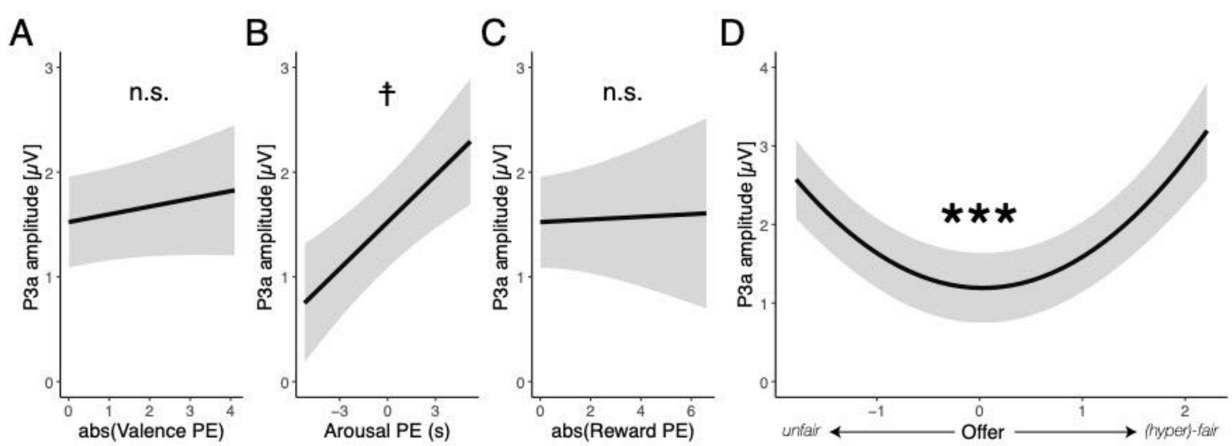


Figure 5. Offer extremity predicts P3a. *A-C. Predicted P3a amplitudes are shown for all three types of absolute PEs (absolute valence, arousal, and reward). Negative values reflect less surprise for all three PEs. D. Predicted P3a amplitudes are plotted by z-scored offers. Analyses*

*reflect regression models and shaded areas reflect ±1 S.E. ***P<.001, **P<.01, *P<.05, †P=.06*

P3b reflects unsigned valence prediction errors and uniquely predicts choice

Finally, we examined whether the last candidate ERP signal, the P3b, uniquely indexes one of the three prediction errors. The P3b occurs after the P3a and is known to reflect processing of unsigned surprising events (Mars et al., 2008) and scales with learning to predict future behavior (Fischer & Ullsperger, 2013). We tested whether any of the unsigned (absolute) PEs are uniquely expressed in the P3b (using the same analysis described above). Results show that absolute valence PEs robustly predict P3b amplitudes ($\beta = 0.34 \pm 0.12, t = .62, P = 0.005$), while arousal PEs ($\beta = -0.03 \pm 0.07, t = -.45, P = 0.65$) and absolute reward PEs ($\beta = 0.23 \pm 0.15, t = 1.52, P = 0.14$) do not (Fig. 5). After establishing that each candidate ERP signal preferentially indexes different PE types, we tested whether any of the three ERPs predict subsequent choice. Consistent with the account that P3b represents adaptation based on offer information (Nassar et al., 2019), we found that only P3b activity predicts choices to punish (Table 2), an effect that decreases over rounds. This echoes our previous behavioral findings demonstrating that valence PEs predict choices strongly on round 1 and weaken over time. In short, these findings indicate that the neural signature of valence PEs is reflected in P3b amplitudes and changes in EEG activity of this component uniquely predicts choice.

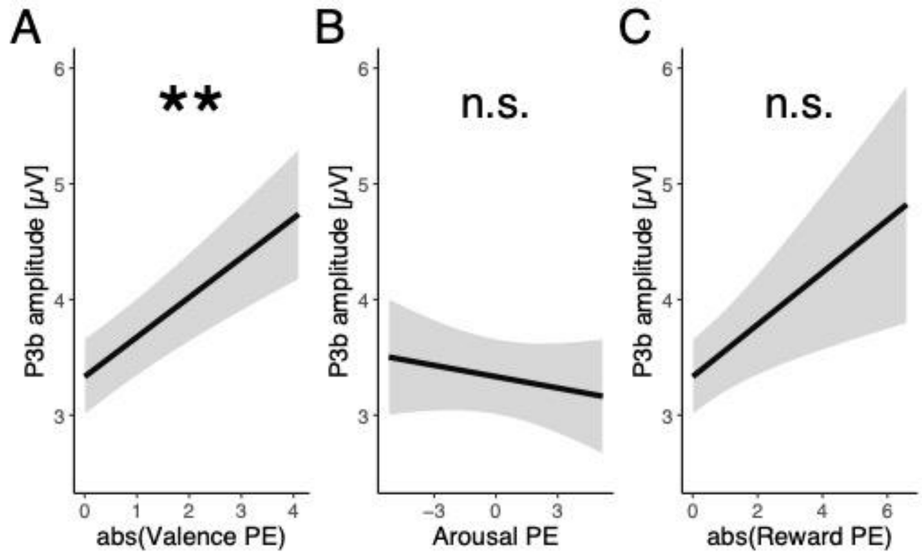


Figure 5. Absolute valence prediction errors (emotion surprise) are integrated in the P3b. *Predicted P3b amplitudes are shown for all three types of absolute PEs (absolute valence, arousal, and reward). Negative values reflect less surprise for all three PEs. Analyses reflect regression models and shaded areas reflect ±1 S.E. ***P<.001, **P<.01, *P<.05.*

Table 2. Only P3b predicts choices over rounds

$$Punish_{i,t} \sim \beta_0 + \beta_1 P3b_{i,t} + \beta_2 P3a_{i,t} + \beta_3 FRN_{i,t} + \beta_4 Round_{i,t} + \beta_5 P2_{i,t} + \beta_6 P3b_{i,t}{:}Round_{i,t} + \beta_7 P3a_{i,t}{:}Round_{i,t} + \beta_8 FRN_{i,t}{:}Round_{i,t} + \varepsilon$$

Variable	Estimate (SE)	z	p
Punish			
Intercept	-1.05 (0.28)	-3.78	<.001***
P3b	0.28 (0.10)	3.03	.002**
P3a	0.21 (0.13)	1.53	.13
FRN	-0.07 (0.13)	-0.56	.57
P2	-0.08 (0.06)	-1.48	.14
Round	-0.02 (0.03)	-0.61	.54
P3b×Round	-0.06 (0.03)	-2.19	.03*
P3a ×Round	-0.05 (0.04)	-1.34	.18
FRN×Round	0.06 (0.04)	1.61	.11

Note. P2, FRN, P3a, and P3b are all trial-by-trial amplitudes in response to the offer (see Methods). All variables were scaled but not mean-centered. The model includes P2 as a control for the FRN since the FRN is defined as the N2 amplitude. The model includes subject-specific random intercept. The dataset includes 5,351 observations from 35 participants.

*** p <.001 ** p <.01 * p <.05

Discussion

Emotion plays an indisputable role in learning and decision-making, yet few reinforcement learning models formalize a role for emotions (with some exceptions; Eldar et al., 2016; Rutledge et al., 2014), focusing instead on reward PEs as a central driver of behavior. Previously, we operationalized the violations of emotion expectations demonstrating a functional dissociation such that emotion PEs predict social choices above and beyond reward PEs (Heffner et al., 2021). Here, we examine the behavioral and neural markers of emotion and reward PEs concurrently during a social learning task while recording EEG signals. Our results show a behavioral dissociation between contributions to learning and choice, such that people rely heavily on emotion PEs to rapidly learn and evaluate offers, and only switch to relying on reward PEs once they are more certain about who they are engaging with. We also test three candidate ERPs that are time-locked to offers—the FRN, P3a, and P3b—finding that these reflect reward, arousal, and valence PEs, respectively. The fact that we found that these three ERPs are each indexing different PEs reveals that reward and emotion error signals are processed on different temporal trajectories. Offers are first evaluated relative to reward surprise (i.e., how does this monetary offer differ from what I was expecting?), and only afterwards are violations of emotion expectations incorporated (i.e., I feel more / less or worse / better than expected). Critically, we find that only the ERP associated with valence PEs—the P3b—uniquely predicts subsequent choices, suggesting that emotion PEs are essential for social updating.

By embedding emotion measurements throughout a social learning paradigm, we were able to measure PEs for different dimensions of emotion (Russell, 2003) to explore how different types of expectation violations are important for learning and decision-making. This broadens previous approaches, which either focused on long timescale affect forecasting errors (Wilson & Gilbert, 2005) or contexts which did not involve trial-by-trial learning (Heffner et al., 2021). We demonstrate that emotion PEs are largely resolved by late rounds, suggesting people learn how to anticipate their future emotional reactions, and valence PEs are a better predictor of choices than reward PEs in early rounds. This suggests that early emotion PEs are particularly important for resolving uncertainty about social partners and making consequential social decisions. In contrast to reward-centric accounts (Silver et al., 2021), these results imply that violations of emotional expectations play a particularly privileged role in social decision-making that cannot be accounted for by reward alone.

The neural signatures of reward PEs and feedback processing are well documented (Holroyd & Coles, 2002; Holroyd et al., 2003; Sambrook & Goslin, 2015), but to date no study has tested whether these ERPs—the FRN, P3a, and P3b—track reward or emotion violations. Our approach examined how social feedback is processed in the brain, and we found evidence that all three PEs can be temporally dissociated from one another. Consistent with reinforcement learning frameworks that implicate the FRN as the neural signature of reward PEs (Hauser et al., 2014; Holroyd & Coles, 2002), we find that reward surprise in response to the offer—regardless of its direction (positive or negative)—was uniquely reflected in FRN amplitudes. Although there is existing evidence that emotional faces and words are processed early on (Baum & Abdel Rahman, 2021; Phelps et al., 2006; Willis & Todorov, 2006), our results suggest that offers are initially evaluated by its reward, and not emotion. We find that absolute valence PEs (emotional surprise) are reflected in the late P3b component, which predicts subsequent choice, which is consistent with the theory that the P3b reflects integration of information under consideration for changes in decision policies (Nassar et al., 2019). The fact that we find that these emotionally surprising error signals are reflected in late ERPs is consistent with evidence that emotion modulates attention during late stages of processing (Dennis & Hajcak, 2009; Liu et al., 2012). In short, these results demonstrate a neural and temporal dissociation where reward PEs are indexed separately and earlier than emotion PEs.

Emotions powerfully shape behaviors, yet few reinforcement learning models incorporate emotions into their models. By situating emotions in a reinforcement learning perspective, we demonstrated how emotion PEs play a central role in social learning alongside traditional reward PEs. Using precise temporal measurements of neural activity, we document evidence for early reward PE and late emotion PE processing of social information. Our results suggest that although violations of emotion expectations are incorporated relatively late in the decision evaluation, they are strongly predictive of choice. Together, these results offer more evidence for the separability of emotion and reward PEs at the neural level and the essential role of emotion PEs in guiding social behavior.

Materials and methods

Participants

Participants (N = 35, 19 female, mean age = 20.9 $\pm$ 4.6) received either monetary compensation ($15 per hour) or course credits and provided informed consent in a manner approved by Brown University's Institutional Review Board under protocol 1607001555. A power analysis of the unique effect of valence prediction error on choice in prior work (Heffner et al., 2021) revealed that 18 participants would be sufficient to detect this effect with an alpha of 0.05 and power (beta) of 0.80. Accordingly, we aimed to exceed this and collected a sample of 35 participants, which matches sample sizes of recent EEG studies focusing on the FRN and P300 (Frömer et al., 2021; Nassar et al., 2019).

Task and procedure

Participants played an adapted repeated Ultimatum Game (Cooper & Dutcher, 2011; Güth et al., 1982) that included subjective emotion and reward ratings (Heffner & FeldmanHall, 2022; Heffner et al., 2021). Participants were told they were playing with past participants who gave offers for five rounds conditional on participant’s choices to accept or reject each offer, similar to strategy methods used in economics (Bahry & Wilson, 2006). Unknown to participants, their partner’s offers were generated from one of three normal distributions representing three types of proposers: 1) “bad” proposers gave offers according to a normal distribution with a mean of $1 and SD of 0.50; 2) “neutral” proposers gave $3 on average with a SD of 0.50; and 3) “good” proposers who gave $5 on average with a SD of 0.50. Participants would play with 36 unique partners, 12 of each type and all five offers from each partner were randomly drawn from their respective normal distribution. We used faces from the 74 image MR2 database (Strohminger et al., 2016) to represent partners and 36 faces were pseudo-randomly pulled from this database per participant to have a balanced distribution of images of men and women of European, African, and East Asian ancestry. Subjective affect predictions and experiences were reported using a 500-500 pixel two-dimensional affect grid where the horizontal axis was valence (unpleasant/pleasant feelings) and the vertical axis was arousal (low/high intensity feelings). To familiarize participants with this affect grid, participants completed an emotion classification task prior to the repeated UG. Participants made affect ratings of 20 canonical emotion words (for example, angry, sad, and surprised) on the grid, twice for each word, in a randomized order. Training participants to interpret this subjective affect grid has shown strong convergent validity with other approaches for emotion ratings (Russell et al., 1989).

The UG consisted of 15 separate screens for each trial, with responses required from participants for only 4 screens. After a fixation cross of 500ms (same timing for all fixation crosses) participants are shown a picture of their partner for 1000ms, followed by another fixation cross. Next, participants are given cues for reward predictions ($?) or emotion predictions (E?) which last 1000ms each; these cues indicate that participants will be making the required response on the next screen. Reward predictions (how much participants expect your partner to offer) are reported on a visual analog scale ($0 - $10) and participants have unlimited time to give this report. Emotion predictions (how participants expect to feel after the offer) are reported on the valence-arousal grid and participants are required to answer within 5s. The order of predictions was counterbalanced and separated by fixation crosses. Following predictions, the offer for the round was given for 2000ms in dollar and cent format (e.g., $2.37) and followed by a fixation cross. Participants were then given an emotion experience rating cue (E) for 1000ms, which indicated that they would be rating how they felt about the offer using the affect-grid. After reporting their emotional experience, which was required within 5s, and a fixation cross, participants were given a choice cue (C) for 1000ms indicating they would be making their choice. The choices to accept or reject were given as capital letters on the screen (e.g., [A] [R]) and the order of these options was counterbalanced. Participants' task was to make their decision, by pressing a response key with their left index and middle finger for the left or right option on the screen respectively. When participants were matched with a new partner, they were presented with a waiting screen for 1-4s before starting the next trial.

The UG comprised nine blocks of 4 partners each (20 trials per block), with self-paced rests between blocks for a total of 180 trials. Participants additionally completed pre-UG and post-UG likability ratings for all 36 partners on a visual analog scale (0 – 10, increments of .01). The entire experiment was programmed in Matlab (The MathWorks, Inc.) using the Psychtoolbox-3 package and included stimulus presentation, event and response logging. A standard computer mouse and keyboard were used for response registration.

Prior to experiment, participants filled in demographic and personality questionnaires: the 20-item Toronto Alexithymia Scale (Bagby et al., 1994) and the Temporal Experience of Pleasure Scale (Gard et al., 2006). These measures were registered as potential control variables and for other purposes not addressed here. Participants were then seated in a shielded EEG cabin, where the experiment including EEG recording was conducted. Prior to completing the full emotion classification and UG task, participants performed five and three practice trials respectively.

Psychophysiological recording and processing

EEG was recorded using a BrainVision recorder software (Brain Products, München, Germany) from 64 Ag/AgCl electrodes mounted in an electrode cap (ECI Inc.). Data was collected using Cz as a reference channel at a sampling rate of 500 Hz and re-referenced to the grand mean for analysis. Electrodes below the eyes (IO1, IO2) and at the outer canthi (LO1, LO2) recorded vertical and horizonal ocular activity. At the end of the experiment we recorded 20 trials of each prototypical eye movement (up, down, left, and right) for offline ocular artifact correction. We kept electrode impedance below 10 kΩ.

EEG data were processed using Matlab (The MathWorks Inc.) using the EEGlab toolbox (Delorme & Makeig, 2004) as previously described (Frömer et al., 2021) and included the following steps: (1) re-calculated the average reference and retrieved the Cz channel, (2) removal of blink and eye movement artifacts using BESA (Ille et al., 2002), (3) bandpass filtering [.1 – 40 Hz], (3), (4) epoch the ongoing EEG from -200 to 800ms relative to offer onset, (5) removal of segments containing artifacts, based on values exceeding ±150 μV and gradients larger than 50 μV between two adjacent sampling points. Baselines were corrected to the 200ms pre-stimulus interval (offer onset).

To define the time windows for single-trial analyses of FRN, P3a and P3b amplitudes, we first determined the average subject-wise peak latencies at FCz, FCz, and Pz, respectively. Accordingly, the FRN was quantified on single trials as the average voltage within an interval from 315 to 415ms after offer onset across all electrodes within a fronto-central region of interest including F3, Fz, F4, FC3,FCz, FC4, C3, Cz, C4 (Yeung & Sanfey, 2004). To control for P2 effects on the FRN, the P2 amplitude was also extracted within each trial as the average voltage between 199-299ms across fronto-central electrodes F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2. P3a amplitude was quantified on single trials as the average voltage within a 363-463ms interval post-offer across fronto-central electrodes F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2 (Frömer et al., 2021). P3b amplitude was quantified on single trials as the average voltage within a 530-630ms interval post-offer within a parietally-focused region of interest including CP1, CPz, CP2, P1, Pz, P2, PO3, POz, PO4 (Frömer et al., 2021).

Analyses

Inspection of the behavioral data identified four trials in which impossible affect ratings were given (valence or arousal ratings outside of the 500 by 500-pixel grid) and these data were excluded from relevant analyses. Reward PEs (RPE) was determined as the offer minus the reward prediction given by participants Valence and Arousal PEs were determined similarly: the affective experience participants reported upon receiving the offer minus participant's affective prediction for how they would feel after the offer. Prior to analyses, reward, valence, and arousal PEs were standardized but not mean centered, as zero represents a meaningful value on these scales (predicted and actual experiences are the same).

References

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1).

Bahry, D. L., & Wilson, R. K. (2006). Confusion or fairness in the field? Rejections in the ultimatum game under the strategy method. *Journal of Economic Behavior & Organization*, *60*(1), 37-54. https://EconPapers.repec.org/RePEc:eee:jeborg:v:60:y:2006:i:1:p:37-54

Baum, J., & Abdel Rahman, R. (2021). Emotional news affects social judgments independent of perceived media credibility. *Social Cognitive and Affective Neuroscience*, *16*(3), 280-291. https://doi.org/10.1093/scan/nsaa164

Collins, A. G. E., & Shenhav, A. (2022). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, *47*(1), 104-118. https://doi.org/10.1038/s41386-021-01126-y

Cooper, D. J., & Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: a meta-study. *Experimental Economics*, *14*(4), 519-546. https://doi.org/10.1007/s10683-011-9280-x

Daw, Nathaniel D., Gershman, Samuel J., Seymour, B., Dayan, P., & Dolan, Raymond J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204-1215. https://doi.org/https://doi.org/10.1016/j.neuron.2011.02.027

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*, *134*(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dennis, T. A., & Hajcak, G. (2009). The late positive potential: a neurophysiological marker for emotion regulation in children. *Journal of child psychology and psychiatry, and allied disciplines*, *50*(11), 1373-1383. https://doi.org/10.1111/j.1469-7610.2009.02168.x

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, *20*(1), 15-24. https://doi.org/10.1016/j.tics.2015.07.010

FeldmanHall, O., Raio, C. M., Kubota, J. T., Seiler, M. G., & Phelps, E. A. (2015). The Effects of Social Context and Acute Stress on Decision Making Under Uncertainty. *Psychol Sci*, *26*(12), 1918-1926. https://doi.org/10.1177/0956797615605807

Fischer, Adrian G., & Ullsperger, M. (2013). Real and Fictive Outcomes Are Processed Differently but Converge on a Common Adaptive Mechanism. *Neuron*, *79*(6), 1243-1255. https://doi.org/https://doi.org/10.1016/j.neuron.2013.07.006

Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cognition and Emotion*, *1*(2), 115-143. https://doi.org/10.1080/02699938708408043

Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). Response-based outcome predictions and confidence regulate feedback processing and learning. *eLife*, *10*, e62825. https://doi.org/10.7554/eLife.62825

Gard, D. E., Gard, M. G., Kring, A. M., & John, O. P. (2006). Anticipatory and consummatory components of the experience of pleasure: A scale development study. *Journal of Research in Personality*, *40*(6), 1086-1102. https://doi.org/10.1016/j.jrp.2005.11.001

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388. https://doi.org/https://doi.org/10.1016/0167-2681(82)90011-7

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, *42*(2), 161-170. https://doi.org/https://doi.org/10.1111/j.1469-8986.2005.00278.x

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*, 159-168. https://doi.org/https://doi.org/10.1016/j.neuroimage.2013.08.028

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(11), 4178-4187. https://doi.org/10.1523/JNEUROSCI.4652-10.2011

Heffner, J., & FeldmanHall, O. (2022). A probabilistic map of emotional experiences during competitive social interactions. *Nature Communications*, *13*(1), 1718. https://doi.org/10.1038/s41467-022-29372-8

Heffner, J., Son, J.-Y., & FeldmanHall, O. (2021). Emotion prediction errors guide socially adaptive behaviour. *Nature Human Behaviour*, *5*(10), 1391-1401. https://doi.org/10.1038/s41562-021-01213-6

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*, *167*, 91-106. https://doi.org/https://doi.org/10.1016/j.cognition.2017.03.006

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev*, *109*(4), 679-709. https://doi.org/10.1037/0033-295x.109.4.679

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, *14*(18), 2481-2484.

Ille, N., Berg, P., & Scherg, M. (2002). Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. *J Clin Neurophysiol*, *19*(2), 113-124. https://doi.org/10.1097/00004691-200203000-00002

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78-83. https://doi.org/10.1126/science.1108062

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

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799-823. https://doi.org/10.1146/annurev-psych-010213-115043

Liu, Y., Huang, H., McGinnis-Deweese, M., Keil, A., & Ding, M. (2012). Neural substrate of the late positive potential in emotional processing. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *32*(42), 14563-14572. https://doi.org/10.1523/JNEUROSCI.3109-12.2012

Lowe, R. (2011). The Feeling of Action Tendencies: On the Emotional Regulation of Goal-Directed Behavior [Review]. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00346

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise. *The Journal of Neuroscience*, *28*(47), 12539. https://doi.org/10.1523/JNEUROSCI.2925-08.2008

Martin, L. N., & Delgado, M. R. (2011). The influence of emotion regulation on decision-making under risk. *J Cogn Neurosci*, *23*(9), 2569-2581. https://doi.org/10.1162/jocn.2011.21618

Nassar, M. R., Bruckner, R., & Frank, M. J. (2019). Statistical context dictates the relationship between feedback-related EEG signals and learning. *eLife*, *8*, e46975. https://doi.org/10.7554/eLife.46975

Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response [https://doi.org/10.1111/j.1469-8986.2010.01057.x]. *Psychophysiology*, *48*(2), 162-175. https://doi.org/https://doi.org/10.1111/j.1469-8986.2010.01057.x

Niv, Y., & Montague, P. R. (2009). Chapter 22 - Theoretical and Empirical Studies of Learning. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics* (pp. 331-351). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-374176-9.00022-1

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042-1045. https://doi.org/10.1038/nature05051

Phelps, E. A. (2005). Emotion and Cognition: Insights from Studies of the Human Amygdala. *Annual Review of Psychology*, *57*(1), 27-53. https://doi.org/10.1146/annurev.psych.56.091103.070234

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science*, *17*(4), 292-299. https://doi.org/10.1111/j.1467-9280.2006.01701.x

Polich, J. (2003). Theoretical Overview of P3a and P3b. In J. Polich (Ed.), *Detection of Change: Event-Related Potential and fMRI Findings* (pp. 83-98). Springer US. https://doi.org/10.1007/978-1-4615-0294-4_5

Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife*, *10*, e61077. https://doi.org/10.7554/eLife.61077

Russell, J., Weiss, A., & Mendelsohn, G. (1989). Affect Grid: A Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology*, *57*, 493-502. https://doi.org/10.1037/0022-3514.57.3.493

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145-172. https://doi.org/10.1037/0033-295X.110.1.145

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252-12257. https://doi.org/10.1073/pnas.1407535111

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473-500. https://doi.org/10.1146/annurev.neuro.23.1.473

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/https://doi.org/10.1016/j.artint.2021.103535

Sosa, M., & Giocomo, L. M. (2021). Navigating for reward. *Nature Reviews Neuroscience*, *22*(8), 472-487. https://doi.org/10.1038/s41583-021-00479-z

Stanton, S. J., Reeck, C., Huettel, S. A., & LaBar, K. S. (2014). Effects of induced moods on economic choices. *Judgment and Decision Making*, *9*(2), 167-175.

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods*, *48*(3), 1197-1204. https://doi.org/10.3758/s13428-015-0641-9

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Willis, J., & Todorov, A. (2006). First Impressions: Making up Your Mind after a 100-Ms Exposure to a Face. *Psychological Science*, *17*(7), 592-598. http://www.jstor.org/stable/40064417

Wilson, T. D., & Gilbert, D. T. (2005). Affective Forecasting: Knowing What to Want. *Current Directions in Psychological Science*, *14*(3), 131-134. https://doi.org/10.1111/j.0963-7214.2005.00355.x

Yeung, N., & Sanfey, A. G. (2004). Independent Coding of Reward Magnitude and Valence in the Human Brain. *The Journal of Neuroscience*, *24*(28), 6258. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Chapter 6

General Discussion

Summary of results

The work in this dissertation provides an empirical account of emotion's role in decision-making, focusing on the role of experienced emotion and the violation of emotion expectations. Because emotion and affect are, by definition, subjective experiences, it has been an enduring challenge to precisely measure them. Any account of emotion's role in social choices must include an acknowledgment of a key component that most researchers agree are common to emotional experiences—affect (FeldmanHall & Heffner, RR American Psychologist). To build out a generalizable framework for studying emotions, the experiments in this thesis combined a low-dimensional, descriptive map of core affect with popular social choice paradigms. We argue that the affect grid in particular provides a strong foundation for measuring emotion. This tool mitigates many of the concerns over measurement (thus sidestepping the debates that have plagued the field for decades) to offer a precise and mathematical approach for measuring affect, that in conjunction with behavioral assays, can be used to infer the nature of emotions in ways that allow results to be generalized.

In Chapter 1, we combined several machine learning algorithms with a high-resolution affect grid to create a probabilistic map of emotion that could be used to classify which specific emotions are experienced in a host of different economic games involving competitive decisions (punishing and defecting). While prior accounts predicted anger as the dominant emotion associated with punishment, our data-driven approach found a diverse array of other negative emotions, such as disappointment or sadness, were far more likely to motivate people to punish another for behaving unfairly. We showed how even basic emotions have large affective variability and that the structure of emotions is organized in a continuous manner along valence and arousal dimensions. This success offered a proof-of-principle for the usefulness of combining a low-dimensional representation of affect with behavioral economic games to precisely test theoretical accounts of emotion's role in social decision-making.

In Chapter 2, we examined the role of emotions in real-world decisions by testing the efficacy of two emotional public health messages, one based on threat and the other on prosocial appeals, during the beginning of the COVID-19 pandemic. We found that both the prosocial and threat interventions were equally effective in increasing participant's willingness to self-isolate, but that they achieved compliancy through different mechanisms. When accounting for the degree of emotional engagement with the public health message (indexed through valence-arousal ratings), results showed that stronger valence and arousal responses to the prosocial message—but not threat—increased willingness to self-isolate. While successful prosocial messages depended on strong, positive emotional engagement, threat messages were similarly effective regardless of the strength of emotional response. Furthermore, the observed heterogeneity of emotional engagement with these messages suggests an ongoing challenge in crafting the emotional language used to influence public attitudes and behaviors. These results demonstrated how our generalizable framework can illuminate the emotional mechanisms behind consequential, real-world decision-making.

In Chapter 3, we developed a novel framework for mathematically computing violations of emotion expectations—termed emotion prediction errors (PEs)—using the high-resolution affect grid. We tested how emotion PEs, compared to reward PEs, drive social decision-making. We demonstrated that both valence and arousal have independent and opposite influences on choices to punish, such that negative valence PEs (more unpleasant than expected) and positive arousal

PEs (more intense feelings than expected) both increased likelihood of punishing another for a moral transgression. Negative reward PEs (less money than expected) also motivated choices to punish, but less so than valence PEs. Furthermore, when examining the temporal dynamics of how emotion unfolds, we uncovered a strong relationship between emotion and choice: early negative emotional reactions (often under one second) predicted people's future punitive choices, and these punitive choices resulted in a rapid positive boost afterwards. To understand whether there is a functional dissociation between emotion and reward PEs, we interrogated how emotion and reward PEs are used in populations characterized by impairments in both reward and emotion processing. Results demonstrated selective impairments in emotion PEs but intact reward PEs when making decisions about punishing norm violators; a pattern that suggests emotion PEs are central to health and adaptive social decision-making.

In Chapter 4, we extended our prior work on emotion PEs to measure how they contribute to trial-by-trial learning and test whether they can be dissociated from reward PEs at the neural level. To answer these questions, participants played a repeated Ultimatum Game while undergoing EEG. We tested three event-related potentials (ERPs) known to be involved in feedback processing as potential candidate signatures of emotion or reward PEs. Our results demonstrated a neural dissociation between reward PEs, indexed by the FRN, and emotion PEs, indexed by both subcomponents of the P300. Behavioral results confirmed our neural separability as we again found that valence PEs outperformed reward PEs in predicting choice and the neural index of valence PEs (P3b) was the only ERP to predict future choices. Additionally, these results demonstrated a specific temporal ordering of the processing of social feedback: offers were first evaluated relative to their reward surprise before the integration of violations of emotion expectations. Together, these findings document the first neural dissociation between reward and emotion PEs and offer insight into how they are incorporated across time to predict choices.

Implications

Together, these chapters outline two major ways emotions enter the decision-making process—experienced and expected emotions—and suggest that models of emotion's role in decision-making ought to focus on the affective qualities (valence and arousal) rather than the qualitative state itself (e.g., anger, sadness, or happiness). For example, predominant views of emotion propose that specific emotions are more likely to trigger specific behaviors (i.e., action tendencies; Frijda, 1986), such as feelings of anger increasing the probability of aggression or punitive responses (Pillutla & Murnighan, 1996). It is further argued that this one-to-one mapping of emotion and behavior highlights an evolutionary adaptive response (Ekman et al., 1987) and the treatment of emotion in the social decision-making literature has reflected this perspective. Our success in predicting decisions, or efficacy of public health messaging, according to valence and arousal suggests that these affective qualities may be the critical component driving choices.

The other major way emotions affect the decision-making process is through violations of emotion expectations. Rather than considering the emotional experience in isolation, our results suggest that the models of emotion's role in decision-making must also incorporate emotion prediction errors as a key factor in learning and evaluation of actions. A popular theory of value-based decision-making is built on the notion of a common currency (Levy & Glimcher, 2012) or the idea that the brain must encode subjective values of different types of rewards (e.g., drops of juice, money, etc.) on a single dimension for comparison. Our behavioral and neural results are compatible with this theory and suggest that reward and emotion processes are integrated to produce adaptive behavior. The fact that reward PEs are first expressed in EEG activity before the incorporating of emotion PEs suggests that emotion's role may be to modify the subjective value of incoming rewards according to expected emotions.

Limitations

Of course, no one tool can overcome all the challenges associated with measuring emotion or affect given its subjective and heterogenous nature. Arguably, there will never be a gold standard measure of emotion as experiential (i.e., consciously reportable subjective feelings), physiological (e.g., skin conductance, fMRI), and behavioral measures all capture different slices of the emotional experience (Mauss & Robinson, 2009). While the relevance of valence and arousal to emotion seems undeniable, we readily admit, as others have argued before us, that the world of emotions cannot be boiled down to a two-dimensional space (Fontaine et al., 2007). The affect grid is only capable of capturing a low-dimensional, descriptive map of consciously accessible subjective feelings, which is just one of many ingredients that comprise a complex emotional experience. For example, without additional metrics the affect grid falls short of being able to distinguish between emotional experiences which occupy similar places in the valence-arousal spectrum (e.g., anger and disgust). Prior research has shown that the affective primitives valence and arousal are sufficient for capturing about half of the variance associated with emotion classifications, but adding more rich representational spaces (e.g., 38 appraisal measurements) significantly increases classification accuracy (Skerry & Saxe, 2015). The affect grid attempts to balance speed and parsimony with only two dimensions, but it cannot outperform measures with significantly more degrees of freedom. Future research should examine which, and how many, additional measures would combine with the affect grid to capture a higher dimensional representation of emotions while not being overly onerous to complete in a one-hour experiment.

There is also disagreement about which two-dimensional space best characterizes affect. Some theorists have proposed a 45-degree rotation of the valence-arousal space that yields new dimensions of positive arousal (approach motivation) and negative arousal (avoidance), and there is evidence that these dimensions might better capture neural activation in areas like the nucleus accumbens when participants are anticipating future rewards or losses (Knutson & Greer, 2008). Other theorists have suggested that a three-dimensional space would better capture the human emotional experience (Demekas et al., 2020) and no doubt there will be more proposed in the future. In principle, the affect grid could be modified to fit any two orthogonal, continuous dimensions (one of its many benefits) and future research should examine which mapping is more intuitive for participants. Finally, because this tool exists in a continuous Euclidean space, it is also difficult to parse either dimension into negative and positive subscales, a method that has been used with other measures (Watson et al., 1988) and which would allow for the identification of mixed (simultaneous negative and positive valence) emotional states (Trampe et al., 2015).

Future directions

One strength of combining our measure of affect with decision-making paradigms is the sheer number of open questions that can be answered. Here we outline a few that would provide interesting new avenues for researchers working at the intersection of emotion and decision-making. First, can we test whether scientists can accurately predict changes in affective states over time. Recent theoretical works situates emotions as an active inference process (Barrett, 2017; Smith et al., 2019), positing that emotions emerge through the processing and interpreting of changes in affect (Cunningham et al., 2013). While these models offer proof of principle that affective processes can be formally modeled, to date, there is a lack of empirical evidence supporting these claims. The affect grid would be an ideal tool to precisely and mathematically document changes in affect as participants read a thrilling novel or watch a romantic movie. By supplementing theory with data, we fill in the gaps in our knowledge of how context changes (e.g., scene transitions) or belief inferences (e.g., learning secrets about a character's past) produce and alter affective states (and vice versa). Second, an open question resolves around how people psychologically represent and structure their affect space and how this representation relates to choice. Asymmetric effects of positive and negative experiences are well documented in other domains (Kahneman & Tversky, 1979; Taylor, 1991), which suggests that different affective representations may also influence choice. For example, a shift in affect that crosses a boundary in affect space (i.e., crossing from unpleasant to pleasant, or low to high intensity) may have a stronger influence on the types of choices made compared to the same shift in affect, but within a boundary or quadrant. This idea, which relates to emotion granularity (Barrett, 2004; Demiralp et al., 2012; Feldman, 1995), might explain how affective reference points and individual differences in the representation of affect space bias choice.

One major benefit of the affect grid is that because it is agnostic to any one emotion theory, it can operate in a manner that is untethered from any emotion subfield, thereby circumventing many of the ongoing debates (Adolphs, 2017). Indeed, almost all (if not all) emotion researchers recognize valence and arousal to be relevant to the experience of emotion, which means that it can easily be used in combination with other methods to broaden our understanding of the human emotional experience. For example, researchers interested in understanding the relationship between self-reported feelings and physiological changes in the body could implement ambulatory at-home electroencephalographic sensors (Eldar et al., 2018), or deploy neuroimaging while subjects make affective ratings on the affect grid. This would allow researchers to map self-reported affective states to neurobiological signals to understand whether there is a one-to-one mapping between self-reported measurements and those elicited from the brain and body. Other researchers may want to understand how appraisals or goals change an individual's affective experiences. To capture higher dimensional features of the emotional experience, such as appraisals of safety or control (Cowen & Keltner, 2017; Ortony & Clore, 2015) or goal states (Nelissen et al., 2007), researchers can merge the affect grid with other emotion frameworks to assess their interactive effect. In each test case, the affect grid does not need to be used a siloed manner and its use should not preclude the reliance on other tools or measurements. In fact, by combining multiple measures of emotions which can each capture different features and levels of the emotional experience, we have the potential to revolutionize our understanding of emotions.

Conclusion

Although we are not the first to propose that a variant of the affect grid can help researchers better understand emotion—indeed, a version of the affect grid dates back to the late 1800s (Wundt, 1903)—we are proposing that modernizing the affect grid will make it a powerful tool for creating a flexible yet unified framework to examine the role of emotion in decision-making.

Especially in the age of big data, it is imperative that decision researchers embrace affective measurements that provide rich, continuous observations to serve as a common foundation across tasks that allow for the generalization of results. Without this, a continued nescience in understanding how even low-dimensional affect influences decision-making, will leave us stumbling in the dark as we search for answers to bigger questions.

References

Adolphs, R. (2017). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience*, *12*(1), 24-31. https://doi.org/10.1093/scan/nsw153

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, *87*(2), 266-281. https://doi.org/10.1037/0022-3514.87.2.266

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1-23. https://doi.org/10.1093/scan/nsw154

Cowen, A., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900-E7909. https://doi.org/10.1073/pnas.1702247114

Cunningham, W. A., Dunfield, K. A., & Stillman, P. E. (2013). Emotional States from Affective Dynamics. *Emotion Review*, *5*(4), 344-355. https://doi.org/10.1177/1754073913489749

Demekas, D., Parr, T., & Friston, K. J. (2020). An Investigation of the Free Energy Principle for Emotion Recognition [Review]. *Frontiers in Computational Neuroscience*, *14*. https://doi.org/10.3389/fncom.2020.00030

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Barrett, L. F., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., Deldin, P. J., Gotlib, I. H., & Jonides, J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychol Sci*, *23*(11), 1410-1416. https://doi.org/10.1177/0956797612444903

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). *Universals and cultural differences in the judgments of facial expressions of emotion* [doi:10.1037/0022-3514.53.4.712].

Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of Reward Learning Signals Predicts Mood Fluctuations. *Curr Biol*, *28*(9), 1433-1439.e1437. https://doi.org/10.1016/j.cub.2018.03.038

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*(1), 153-166. https://doi.org/10.1037/0022-3514.69.1.153

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050-1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Frijda, N. H. (1986). *The emotions*. Editions de la Maison des Sciences de l'Homme.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1511), 3771-3786. https://doi.org/10.1098/rstb.2008.0155

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027-1038. https://doi.org/https://doi.org/10.1016/j.conb.2012.06.001

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & emotion*, *23*(2), 209-237. https://doi.org/10.1080/02699930802204677

Nelissen, R. M. A., Dijker, A. J. M., & de Vries, N. K. (2007). Emotions and goals: Assessing relations between values and emotions. *Cognition and Emotion*, *21*(4), 902-911. https://doi.org/10.1080/02699930600861330

Ortony, A., & Clore, G. (2015). Can an appraisal model be compatible with psychological constructionism? In *The psychological construction of emotion.* (pp. 305-333). The Guilford Press.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, *68*(3), 208-224. https://doi.org/10.1006/obhd.1996.0100

Skerry, Amy E., & Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, *25*(15), 1945-1954. https://doi.org/10.1016/j.cub.2015.06.009

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating Emotions: An Active Inference Model of Emotional State Inference and Emotion Concept Learning [Hypothesis and Theory]. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02844

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychol Bull*, *110*(1), 67-85. https://doi.org/10.1037/0033-2909.110.1.67

Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0145450

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol*, *54*(6), 1063-1070. https://doi.org/10.1037//0022-3514.54.6.1063