A computational framework for emotion understanding

by

Sean Dae Houlihan

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2022

Author .....
Department of Brain and Cognitive Sciences
August 25, 2022

Certified by .....
Rebecca Saxe
John W Jarve (1978) Professor, Associate Dean of Science
Thesis Supervisor

Accepted by .....
Mark Harnett
Chairman, Department Committee on Graduate Theses

A computational framework for emotion understanding

by

Sean Dae Houlihan

Submitted to the Department of Brain and Cognitive Sciences
on August 25, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The organizing principle of this thesis is that human emotion understanding reflects a model-based solution to a large class of ill-posed inverse problems. To interpret someone's expression, or predict how that person would react in a future situation, observers reason over a logically- and causally-structured intuitive theory of other minds. For this work, I chose a domain that is perceptually and socially rich, yet highly constrained: a real-life high-stakes televised one-shot prisoner's dilemma.

In the first set of studies, I illustrate that forward predictions play a critical role in emotion understanding. Intuitive hypotheses about what someone is likely to feel guide how observers interpret and reason about expressive behavior. By simulating human causal reasoning as abductive inference over latent emotion representations, a parameter-free Bayesian model captured surprising patterns of social cognition.

In the second set of studies, I formalize emotion prediction as a probabilistic generative model. Mental contents inferred via the inversion of an intuitive theory of mind generate the basis for inferring how others will evaluate, or 'appraise', a situation. The *Inferred Appraisals* model extends inverse planning to simulate how observers infer others' reactions, in the terms of utilities, prediction errors, and counterfactuals on rich social preferences for fairness and reputation. I show that the joint posterior distribution of inferred appraisals provides a powerful method for discovering the latent structure of the human intuitive theory of emotions.

In the third set of studies, I build a stimulus-computable model of emotion understanding. This work emphasizes the importance of testing whether computational models can use emotion-relevant information in service of social cognition. I suggest that building computer systems that approach human-level emotional intelligence requires generative models, where inferred appraisals function as latent causal explanations that link behavior, mental contents, and world states.

Thesis Supervisor: Rebecca Saxe
Title: John W Jarve (1978) Professor, Associate Dean of Science

Acknowledgments

> Still, it would take a whole other book to convey the gratitude I feel when I look back on this project.
>
> — John Koening, *The dictionary of obscure sorrows*

Tracing the thread of friends and mentors that leads to this point makes abundantly clear that there is simply no adequate way to express my gratitude to all those who deserve it or to the level I feel it. This thesis would not exist without Rebecca Saxe. While the work that follows is my doctoral *thesis*, it would be more appropriate to think of it as a *synthesis*, because it so fundamentally reflects the dialectic between Rebecca and me. Rebecca, your unwavering devotion to my development as a scientist is only paralleled by your kindness towards me. When I started graduate school I did not imagine I would ever do a project with you but I was inspired by your ability to find the interface between interesting questions and interesting answers, and was happy to fabricate opportunities to absorb your scientific approach. My admiration has only grown in the intervening years. I am beyond lucky to have had so much time to learn with you.

Josh Tenenbaum, you have shaped the foundations of my scientific philosophy. The way I conceive of understanding, of meaning, and of progress, reflects your influence. I am incredibly grateful for how you have supported and encouraged me as I have navigated epistemology and methodology on the way to finding processes and projects that speak to me. John Gabrieli, you have helped me stay connected to my deepest passions in science. I cannot thank you enough for your unconditional support. You have taught me how to simultaneously aspire to grand ideals and value incremental progress. I have learned much about enthusiastic perseverance from your example. Luke Chang, you always see the best in me and my work, even when I do not.

I've had the fortune to work with many incredible people over the course of this degree. Stefano Anzellotti, thank you for your patience and for being an incredible teacher and friend. It has been a joy to collaborate with Max Kleiman-Weiner and Desmond Ong, this work has greatly benefited from your involvement. My gratitude to the exceptional post-baccalaureate and undergraduate students I have worked with: Brandon Davis, Riana Hoagland, Antonis Michael, Jimmy Capella, Samuel Liburd Jr., Schuyler Gaillard, Annmarie Wang, Selena Feng, and others. Thank you to the past and present members of SaxeLab, CoCoSci, and GabLab, who have helped me prepare talks, given feedback on papers, and been a constructive sounding board for many inchoate ideas.

I did not expect to find such dear friends in my cohort. Sarah Schwettmann, Luke Hewitt, and Maddie Cusimano, you have made the daily work fun, the progress possible, and the memories deeply fulfilling. Thanks to Matthias Hofer and Tyler Brooke-Wilson for cultivating a community of play, and Venerable Tenzin Priyadarshi for cultivating a community of practice.

I am grateful to the extraordinary mentors who helped me transition into this line of work from molecular biology: Jud Brewer, Noopur Amin, Sue Whitfield-Gabrieli, Daniela Kaufer, Adam Engle, Dustin DiPerna, John Churchill, and Barbara Des Rochers. You took a chance on me and went far out of your way to help me acquire skills I needed. Dan Brown and Gretchen Nelson, when I'm with you, what's most important in life is crystal clear. Dan, you started me on this path. This thesis is dedicated to you.

Teresa Yeh, thank you for the joy you bring into my life. I can't imagine spending these last years with anyone else. To my parents, Mickey and Judy, there are no words that can possible approach how grateful I am. You inspire me to live more fully and love more bravely.

Contents

1 Introduction **9**

1.1 Research strategy 10

1.1.1 Theoretical framework 10

1.1.2 Modeling strategy 13

1.1.3 An experimental domain for progress 14

1.2 Advancing the modeling enterprise 15

1.3 Summary 19

2 Formalizing emotion concepts within a Bayesian model of theory of mind **22**

2.1 Introduction 22

2.2 Situating emotion concepts within an intuitive theory of mind 23

2.3 Specificity and development of emotion inference 25

2.4 Ambiguous perception and precise predictions 27

2.5 Neural representations of fine-grained emotion concepts 29

2.6 Conclusion 30

3 Reasoning about emotions, expressions, and events **31**

3.1 Introduction 31

3.1.1 Spontaneous expressions in a high-stakes social dilemma . . . 37

3.2 Study 1: Observers make nuanced and reliable emotion predictions based on event descriptions 40

3.2.1 Methods 40

3.2.2 Results and Discussion 41
3.2.3 Summary 43
3.3 Study 2: Emotions conveyed by expressions do not discriminate between event contexts 44
3.3.1 Methods 45
3.3.2 Results and Discussion 46
3.3.3 Summary 47
3.4 Study 3: Expressions are interpreted in light of conceptual knowledge 48
3.4.1 Methods 49
3.4.2 Results and Discussion 49
3.4.3 Summary 52
3.5 Study 4: Systematic errors in human causal reasoning about the antecedents of expressions 53
3.5.1 Methods 55
3.5.2 Results 56
3.5.3 Confidence accentuates intuitive theory 58
3.5.4 Collective judgments 59
3.5.5 Summary 59
3.6 Study 5: Abductive inference over latent emotion representations . . 62
3.6.1 Abductive inference model 63
3.6.2 Results 65
3.6.3 Lesion models 67
3.6.4 Summary 70
3.7 General Discussion 71
3.8 Methods 77

4 Generative model of inferred appraisals 83

4.1 Introduction 83
4.2 Background 86
4.3 Inverse planning with social values 90

4.3.1 Modeling inverse planning with social values 92
4.3.2 Comparison to human inverse planning 94
4.3.3 Second-order preferences: players' motive to enhance their reputation 96
4.4 Emotion predictions 99
4.4.1 Human observers' emotion predictions 101
4.4.2 Learning the latent structure of the intuitive theory of emotions 102
4.4.3 Comparing the Inferred Appraisals model to human observers 105
4.4.4 Inverse planning lesion model 107
4.4.5 Social lesion model 109
4.5 Personalizing emotion predictions 111
4.5.1 Simulation of the bias induced by personalizing cues 112
4.6 Conclusion 115
4.7 Methods 118

5 Stimulus-computable emotion understanding 126
5.1 Introduction 126
5.1.1 Related work 128
5.2 Task 1: Attribution of emotions to expressions in context 129
5.2.1 Empirical data 129
5.2.2 Model 130
5.2.3 Results 134
5.2.4 Discussion and Directions 136
5.3 Task 2: Causal inference of antecedent events from expressions 138
5.3.1 Empirical data 139
5.3.2 Model 139
5.3.3 Results 141
5.3.4 Perceptual outcome classification 141
5.3.5 Discussion and Directions 142
5.4 General Discussion 145

5.5 Conclusion 150
5.6 Methods 152

A Appendix to Chapter 3 **156**
A.1 Limitations 156
A.2 Supplementary Analyses 157
A.2.1 Statistics of actual gameplay 157
A.2.2 Similarity structure of emotion judgments 158
A.2.3 PCA of emotion judgments 159
A.2.4 Reliability of emotion judgments 161
A.2.5 Conceptual knowledge affects the interpretation of expressions 163
A.2.6 Human causal reasoning 165
A.2.7 Simulation of collective outcome judgments 166
A.2.8 Abductive inference model 167

B Appendix to Chapter 4 **169**
B.1 Methods 169
B.1.1 Mental content attribution prompts 169
B.1.2 Utilities 170
B.1.3 Planning 171
B.1.4 Appraisals 171
B.2 Personalized priors 172

Chapter 1

Introduction

> Beyond the variegated sensations and the helpful motivations, science has discovered emotionality's deeper purpose: the timeworn mechanisms of emotion allow two human beings to receive the contents of each other's minds.... This silent reverberation between minds is so much a part of us that, like the noiseless machinations of the kidney or the liver, it functions smoothly and continuously without our notice.
>
> — T. Lewis, F. Amini, R. Lannon, *A general theory of love*

This thesis begins with an aspiration and a problem. The mere shift of a gaze, or flicker of a smile, can change what we think someone feels. These shifts are often inconsequential, but can be seismic. What we think about others' emotions shapes our understanding of their values, beliefs, intentions, prior experiences, and physiology, and similarly updates our understanding of the world, informs our self-knowledge, guides our plans, and affects our own emotional experiences.

The computational framework I present aspires to reverse engineer human emotion understanding: to uncover the mechanisms by which we bridge the space between minds. People across many fields share the goal of understanding how that bridge is constructed, when it holds, when it crumbles, and when it takes us somewhere unintended. Over the last one hundred years, psychologists, statisticians, philosophers, and more recently, computer scientists, have tried to distill the essential mechanisms of emotion understanding from the complexity of everyday experience. These efforts

have, on one hand, tended to yield theories that make concrete and specific predictions, but only account for human behavior under narrow experimental conditions. On the other hand, these efforts have yielded theories that are qualitatively consistent with a breadth of behavior, but lack the formalism required to quantitatively emulate human cognition. The problem is therefore how to model the sophistication, variability, and generality of human emotion understanding.

This thesis aims to ground the study of human emotion understanding in formal models and explicit, testable assumptions. The organizing principle of this work is that human emotion understanding reflects a model-based solution to a large class of ill-posed inverse problems. This framing argues that emotion-relevant information, including expression cues and event context, should be studied in terms of their functional roles in a logically- and causally-structured intuitive theory of mind. The present chapter situates the specific contributions of subsequent chapters in the broader research program.

1.1 Research strategy

1.1.1 Theoretical framework

While emotion understanding is often treated as perceptual pattern matching, a growing body of work has convincingly challenged some of the foundational assumptions for this approach (reviewed in Chapter 2). To develop a different theoretical approach, I propose initializing at a very different set of assumptions. Rather than expressive behavior inherently signaling rich diagnostic information about someone's internal states, what if perception of isolated physical expressions largely supplies ambiguous, low-dimensional, and noisy information? To compensate, observers make specific granular emotion predictions based on inferences of how someone interprets (or "appraises") external events in relation to her other mental states (goals, beliefs, moral values, costs, etc.) (de Melo et al., 2012; Gratch & Marsella, 2014). These *inferred appraisals* are abstract representations that function as latent causal expla-

nations: inferred appraisals link expressions, actions, mental contents, and world states in a causally-structured intuitive theory (Figure 1-1). Situating these variables in this common latent space provides a way to quantitatively model how contextual cues and expressive behavior mutually constrain emotion inference.

This framework treats emotion understanding as causal inference over an intuitive theory. An intuitive theory is a mental model that enables commonsense reasoning by connecting concepts in a logically-structured ontology (Carey, 2009; Gerstenberg & Tenenbaum, 2017; Gopnik & Wellman, 2012). In the proposed framework, emotion concepts reflect computations in the space of inferred appraisals. From the context, observers predict what emotions someone is likely to experience based on commonsense reasoning about how she is likely to appraise the situation. From her expression, observers reason about what emotions and appraisals can explain her expressive behavior. In the context of intuitive hypotheses about what experiences someone might be having, otherwise inconsequential shifts in posture or temporal synchrony could become laden with meaning. Thus, constraints from emotion predictions can render expressions conditionally informative: expressions that are ambiguous in one context might become informative in another.

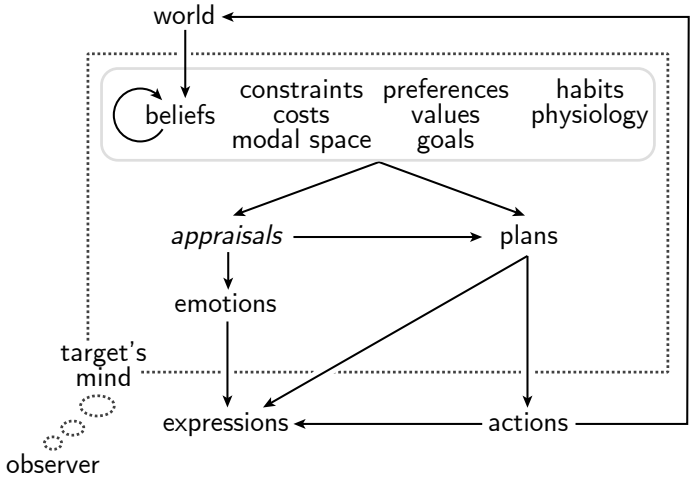

Figure 1-1: **Intuitive theory** (see Chapter 2). The schematic represents a scientific hypothesis about the conceptual structure of emotion understanding. Arrows indicate causal relations. Variables can be inferred via model inversion (e.g. what preferences were likely to have motivated someone's action) and simulated from structured priors. The dimension of time is implicit (variables update and inferences can be made across time). Note that the intuitive theory need not be introspectively accessible.

Highlighting the importance of emotion predictions motivates the effort to model context. Contextual cues affect what emotions we think others are likely to experience and, correspondingly, how we interpret their expressions. Is a medical doctor currently volunteering with a global health NGO likely have the same emotional reaction about being caught in a self-serving lie as the CEO of Exxon? Does the smile of someone handing you a gift mean the same as the identical muscular conformation on the face of someone stealing your wallet? To make progress towards capturing the breadth, flexibility, and nuance of human emotion understanding, we need to model how people use abstractions of contextual information when reasoning about other minds.

Contextual abstractions exist at multiple levels. People readily incorporate explicit event information into their judgments of expressions (Anzellotti et al., 2021; Ong et al., 2015; Skerry & Saxe, 2014). In the absence of explicit contextual cues, e.g. in studies that have people judge emotions based exclusively on expressive behavior, emotion judgments nonetheless reflect contextual priors, including constraints inferred from a task's demand characteristics (Betz et al., 2019; Hoemann et al., 2019) and acquired over a lifetime of experience (Brooks & Freeman, 2018).

Recent work has illustrated surprising ambiguity of isolated expressions (reviewed in Chapters 2, 3, and 5), but the malleability of expression interpretation was noted long before (e.g. Fernberger, 1928; Ruckmick, 1921; Russell, 1994; Wallbott, 1988). The scientific discourse has predominantly revolved around the degree to which context affects emotion understanding (Barrett et al., 2011; Buck, 1994; Cowen & Keltner, 2020; Ekman et al., 1972; Le Mau et al., 2021; Russell & Fehr, 1987). Relatively little attention has been given to developing formal models that can quantitatively predict how contextual information affects emotion understanding. This is due in no small part to the challenge of computing the relevant abstractions from contextual information (Houlihan et al., 2021; Hwang et al., 2021; Lake et al., 2017; Shu et al., 2021). Heider and Simmel highlight this challenge in the opening of their seminal 1944 paper,

> It is true that there have been studies concerning the inference of emotions from gestures or facial change. But most of these leave the reader with a feeling of disappointment and with the conviction that facial 'expressions'—at least as taken by themselves—do not play an important role in the perception of other persons. We are usually referred to the 'importance of the situation'; but what features of the situation are of importance or how the situation influences the perception are problems which are left unanswered.

Nearly eight decades later, their assessment still rings true. How might we make progress towards a principled account of what features of a situation are important and how those features influence the interpretation of perceptual information?

1.1.2 Modeling strategy

It may be possible to learn the necessary representations from vast amounts of data. Advances in deep learning have recently made progress in capturing behavioral patterns in social cognition tasks (Hwang et al., 2021; Liang et al., 2022; Rabinowitz et al., 2018). However, it is a challenge to acquire inductive biases that permit more than surface-level capture of human social reasoning (Stojnic et al., 2022). Richly-structured generative models evidence better generalization and more granular reasoning, even in tasks simple enough for human infants (Shu et al., 2021; Zhi-Xuan et al., 2022).

This thesis adopts a richly-structured modeling approach (for reviews of related work, see Houlihan et al., 2021; Ong et al., 2019; Wu et al., 2021) Endowing models with primitive psychological representations (agents, actions, preferences, beliefs, relations, etc.) affords interpretability, tunability, and inductive constraints. In conjunction with a conducive experimental paradigm, this cognitive structure may enable us to understand how people reason over an intuitive theory of emotions in general.

The overarching research strategy is to leverage experimentally tractable domains to uncover the computational foundations of emotion understanding. Then, on these foundations, build increasingly general models that span more naturalistically unconstrained domains of social cognition. Advancing the modeling enterprise therefore requires a theoretical framework, a modeling strategy, and also an experimental do-

main in which to make modeling progress. In this thesis, I chose a domain that is perceptually and socially rich, yet highly constrained: a real-life high-stakes televised one-shot prisoner's dilemma.

1.1.3 An experimental domain for progress

Every episode of the British gameshow "GoldenBalls" culminates with two contestants playing a dramatic one-shot instantiation of the prisoner's dilemma. Each player is given a choice to "Split" or "Steal" a jackpot (in standard notation, to "`C`ooperate" or "`D`efect", respectively). The game is emotionally evocative by design. Players negotiate with each other in front of a live audience in an attempt to convince the other to make a choice that is financially disadvantageous (to cooperate), then simultaneously reveal their choices.

I generated stimuli from archival footage of the show by artificially separating contextual information about the events players experienced from perceptual information about the players' expressions. The event that a player experienced is defined by the rules of the gameshow, the size of the jackpot, which actions the two players chose, and the resulting financial payoff. I created a 5-second video of each player's emotional expression by splicing together footage from the moments surrounding the climactic reveal.

This experimental paradigm is particularly well-positioned to make progress towards learning the computational basis of emotion understanding. First, segregating perceptual expression information from contextual information allows the overall intuitive theory (Figure 1-1) to be broken down into modules, which can be studied independently and in interaction (Figure 1-2). Second, the event context affords experimental control. Economic games like the prisoner's dilemma associate social interactions with quantitative actions and outcomes. This is conducive to building and fitting models since the context can be experimentally manipulated and separate events can be parametrically related. Third, the events are salient and social. The GoldenBalls game involves high-stakes social coordination, trust, betrayal, equity, and public reputation. Since the goal is to eventually capture the sophistication, nu-

ance, breadth, and flexibility of the intuitive theory of emotions, it is important that the experimental domain elicit rich mental state inferences.

Fourth, the expressions were spontaneously produced by real people reacting to events they actually experienced. In the interest of experimental control, behavioral paradigms commonly use posed expressions or fabricated events. Expression interpretation is sensitive to experimental demand characteristics (Betz et al., 2019; Hoemann et al., 2019). If observers believe that an expression is posed, they are likely to relate to the expression as a symbol with communicative intent. Rather than treating the expression as information about latent emotional experience of the target, observers might be biased by inferences of what the experimenter or the target intends the expression to communicate. This may be a factor in paradigms that employ canonical expressions of basic emotions, muscular conformations enacted by computer-generated avatars, or expressions of people acting out a scene. Similarly, while it is easy for observers to discount posed expressions that seem incongruent with the event context, I expect that observers engage different cognitive processes when they believe that they are seeing the real expression someone spontaneously produced in reaction to an event actually experienced.

1.2 Advancing the modeling enterprise

Chapter 2 outlines the theoretical framework for computationally modeling human emotion understanding. I argue that emotion understanding should be thought of as causal reasoning over a richly-structured intuitive theory. This theory-based view contrasts with alternative "perceptual pattern recognition" views of emotion understanding. Drawing on constructivist theories of emotion, appraisal theories, and inverse planning, I propose how emotion understanding can be formally modeled as a hierarchical Bayesian theory of mind. The framework points to inferred appraisals as the core latent space of the intuitive theory of emotions. Inferred appraisals provide a way to formally model how forward predictions from context and inverse inferences from expressions mutually constrain emotion understanding. I point out evidence

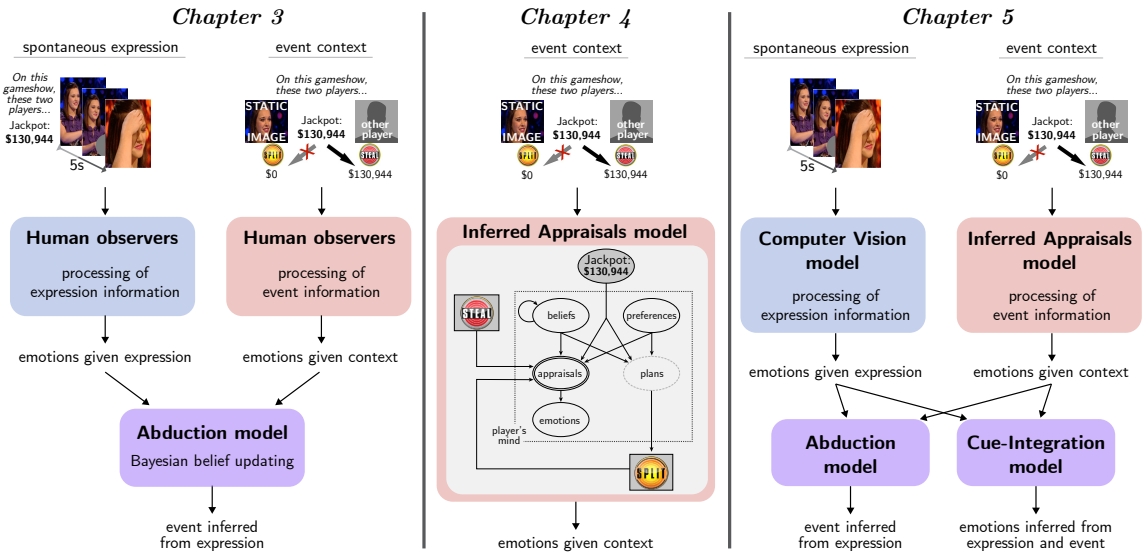

Figure 1-2: **Contrastive chapter summary.** The computational framework proposed in Chapter 2 is empirically developed in Chapters 3, 4, and 5. Chapter 3 characterizes what information observers abstract from expressions and from context. A computational model captures how observers use context to causally reason about expressions. Chapter 4 formalizes how observers predict others' emotions based on contextual information. It illustrates that a richly-structured generative model of inferred appraisals can learn the computational basis of the intuitive theory of emotions. Chapter 5 integrates the preceding chapters with computer vision models to build a stimulus-computable model of social cognition is the chosen experimental domain.

consistent with this view and neural mechanisms that might support the ability to reason about others' emotions.

Chapters 3, 4, and 5 empirically develop this framework in the chosen experimental domain. I make use of the modular GoldenBalls paradigm to study the cognitive mechanisms of emotion understanding (Figure 1-2). Each chapter systematically varies what information people observed, and what judgments they were asked to report. The datasets used in each chapter are summarized in Table 1.1.

Chapter 3 demonstrates how observers reason about expressions by using emotion predictions abstracted from contextual cues. I used a combination of behavioral experiments and a computational model to study how observers interpret the dynamic expressions players spontaneously produced in the GoldenBalls game. The results support key theoretical assertions of Chapter 2: (i) in isolation, even perceptually rich

expressions can be strikingly ambiguous; (ii) observers use intuitive hypotheses about what someone is likely to experience to causally reason about expressive behavior in context; (iii) expression interpretation is better explained by intuitive theory-based causal reasoning than by perceptual pattern matching.

I found that human observers were remarkably poor at recovering what events players in the GoldenBalls game were reacting to, based on their expressions. Beyond being simply inaccurate, people's causal reasoning showed systematic model-based patterns of errors. To reveal the underlying mechanism, I tested if people's inferences could be explained by their latent representations of emotion. I show that people have a reliable, but not necessarily accurate, intuitive theory of the emotions others are likely to experience in hypothetical situations. In certain situations, the emotions people were expected to experience were dramatically different than the emotions they appeared to experience. Observers' interpretations of expressions were a function of conceptual knowledge about what events were possible: inducing a prior over events that shaped the space of predicted emotions biased how observers interpreted expressions.

Integrating these behavioral results, I show that latent emotion representations can explain people's reasoning about the unseen causes of observed expressions. A hierarchical Bayesian model simulated human causal reasoning by comparing the emotions that were inferred from people's expressions against the emotions they were predicted to experience in each situation. Abductive inference over this model provides a close, parameter-free fit to human judgments. This work suggests that humans interpret others' expressions in the context of emotion predictions generated by a causally-structured mental model of other minds.

Chapter 4 formalizes emotion prediction as a probabilistic generative model. The preceding chapter supported the central role of contextually-based predictions in emotion understanding, but did not address how observers compute emotion predictions from context. I now show that inferred appraisals, the core of the computational framework proposed in Chapter 2, naturally capture the fine-grained structure of

emotion predictions.

Humans readily make forward predictions of what emotions others are likely to feel based on information about the events they experience. Observers' systematic predictions about other people's reactions to events reflect an intuitive theory about the causal structure of emotions. Building on inverse planning models, I extend Bayesian theory of mind into the domain of emotions. The Inferred Appraisals model simulates how people use event knowledge to reason about others' nuanced emotional experiences. The model predicts people's judgments of both diverse and nuanced emotions from situation-computable variables. In addition to capturing the effects of explicit event context, the Inferred Appraisals model is sensitive to how person-knowledge, a type of abstract contextual information, biases emotion judgments.

This chapter advances the tractability of using event structure in formal models of human emotion understanding. The richly-structured generative formulation provides a basis for discovering the representations and computations underlying observers' emotion predictions. The latent space of inferred appraisals enables the computational structure of the intuitive theory of emotions to be learned directly from emotion judgments. Finally, the model points to a strategy for computing emotion-relevant abstractions from context, and suggests what primitive representations will be generally required.

Chapter 5 builds a stimulus-computable model of emotion understanding. The modular experimental paradigm allowed me to model how observers combine emotion representations abstracted from expressions and context (Abductive Inference model, Chapter 3), and then model how observers abstract emotion representations from context (Inferred Appraisals model, Chapter 4). I now use (i) off-the-shelf computer vision models to abstract emotion representations from expressions, (ii) the Inferred Appraisals model to abstract emotion representations from context, and (iii) the Abductive Inference model to simulate causal reasoning about expressions (the target behavior from Chapter 3) without human involvement. Additionally, I use the same emotion representations to simulate a separate social cognitive task: inference of

emotions when expressions and context are jointly observed.

This work offers a benchmark of how well stimulus-computable models can capture human behavior and identifies the most important areas for development. I found that performance on both tasks was limited by the computer vision models, whereas the Inferred Appraisals model supported near human-level performance. Importantly, the computer vision models could be coerced to yield relatively good fits to the target emotion representations (human emotion judgments of isolated expressions). Both social cognition tasks amplified the errors of the computer vision models. In other words, objective functions should target social understanding and explanation: within-domain losses (e.g. on emotion judgments of expressions) can belie cognitive relevance. This emphasizes the importance of testing whether models can *use* emotion-relevant information in service of social cognition. Based on these results, I highlight specific but interconnected challenges in processing expressions and event context. Finally, I suggest directions for building computer systems that approach the emotional intelligence of humans.

1.3 Summary

This thesis develops a framework for computationally recapitulating human emotion understanding. I argue that expression cues and contextual information mutually constrain inference over an intuitive theory of mind. Emotion concepts reflect computations in the space of inferred appraisals. Inferred appraisals function as latent causal explanations that link others' expressions, actions, preferences, beliefs, costs, and world states across time.

This work aims to demonstrate how theory, formal models, and experimentally tractable domains can make progress towards quantitatively capturing social cognition and building machines with human-like emotional intelligence. While the GoldenBalls paradigm is the present domain, the intellectual progress is not specific to prisoner's dilemma, gameshows, the expressions of these particular players, or the population of observers who contributed empirical data. The main contribution

of this thesis is a framework for learning the computational foundations of the human intuitive theory of emotions.

In my view, advancing the broader research program requires generative models of how humans causally reason about the variety of perceptually-available expression cues and the dauntingly unconstrained domain of contextual cues that we naturalistically encounter. How to learn and build generative probabilistic programs that capture the breadth and nuance of emotion understanding remains an open problem. So, in a way, this thesis ends where it began: with an aspiration and a problem. But also, hopefully, a deeper understanding of the reverberations between minds.

- A version of Chapter 2 was published as: Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology.*

- A version of Chapter 3 was published as: Houlihan, S. D., Ong, D., Cusimano, M., & Saxe, R. (2022). Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory of mind. *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Data Set	Dependent Variables (DV)	Independent Variables (IV)	Veridical IV	Contextual Frame	Chap
(i)	$\boldsymbol{e}^{adj}$	a_1, a_2, *pot*	✓	GoldenBalls	3, 5
(ii)	$\boldsymbol{e}^{adj}$	x	N/A	no-context	3, 5
(iii)	$\boldsymbol{e}^{adj}$	x, *pot*	✓	GoldenBalls	3, 5
(iv-a)	a_1^{self}, a_2^{self}	x, *pot*	✓	GoldenBalls	3, 5
(iv-b)	a_1^{others}, a_2^{others}	x, *pot*	✓	GoldenBalls	3, 5
(v)	$\boldsymbol{\omega}^{base}$, π_{a_2}	a_1, *pot*	✗	*AnonymousGame*	4, 5
(vi)	$\boldsymbol{\omega}^{base}$, $\boldsymbol{\omega}^{repu}$, π_{a_2}	a_1, *pot*, *SpecificPlayer*	✗	GoldenBalls	4, 5
(vii)	$\boldsymbol{e}^{noun}$	a_1, a_2, *pot*	✗	GoldenBalls	4
(viii)	$\boldsymbol{e}^{noun}$	a_1, a_2, *pot*, *SpecificPlayer*	✗	GoldenBalls	4
(ix)	$\boldsymbol{e}^{adj}$	x, a_1, a_2, *pot*	✓	GoldenBalls	5
(x)	$\boldsymbol{f}^{Azure}$	x	N/A	no-context	5
(xi)	$\boldsymbol{f}^{Rekognition}$	x	N/A	no-context	5

Table 1.1: **Empirical Data Sets (i-ix).** Human participants made judgments of the Dependent Variables (DV) based the Independent Variables (IV). When multiple IV are given, the Veridical IV column indicates whether the IV reflect true values and pairing (based on the GoldenBalls footage) or have been manipulated. The Contextual Frame indicates what background information observers were given during the training phase of the experiment: observers were told the rules of GoldenBalls and watched an example of two players negotiating in the lead up to revealing their decisions (GoldenBalls)†; observers were told that the game was anonymous, rather than a gameshow with a negotiation between the players (*AnonymousGame*); observers were given no information about the provenance of the expressions (no-context). The stimulus variables are: the 5-second silent expression video showing the dynamic expression spontaneously produced by the focal player (x); the focal player's action ($a_1 \in \{\texttt{C}, \texttt{D}\}$); the opponent's action ($a_2 \in \{\texttt{C}, \texttt{D}\}$); the size of the jackpot (*pot*). In Chapter 4, observers were also given a description of a player's occupation (*SpecificPlayer*). The DV include: the intensities of 20 emotions, rated on continuous scales ($\boldsymbol{e}^{adj}$ - adjective label set, $\boldsymbol{e}^{noun}$ - noun label set); each player's action, rated on a 3-point confidence scale (a_1^{self}, a_2^{self}); judgments about what inferences other participants are likely to make about the players' actions (a_1^{others}, a_2^{others}); the focal player's first-order and second-order preference weights, rated on continuous scales ($\boldsymbol{\omega}^{base}$ and $\boldsymbol{\omega}^{repu}$, respectively); what decision the focal player thinks the opponent will make, rated on a 3 point confidence scale (π_{a_2}). Participants were always shown a static photo of the focal player, which was taken before the players revealed their actions.

Computer Vision Data Sets (x,xi). Microsoft Azure Emotion Detector and Amazon Rekognition returned confidence rating time series of 8 emotions ($\boldsymbol{f}^{Azure}$ and $\boldsymbol{f}^{Rekognition}$) based on an expression video.

† The GoldenBalls contextual frame is referred to in Chapter 3 as "broad-context", in Chapter 4 as "Public Game", and in Chapter 5 as parameterization "c".

Chapter 2

Formalizing emotion concepts within a Bayesian model of theory of mind

> "I'd rather write an encyclopedia about common emotions," he admitted. "From *A* for 'Anxiety about picking up hitchhikers' to *E* for 'Early risers' smugness' through to *Z* for 'Zealous toe concealment, or the fear that the sight of your feet might destroy someone's love for you.' "
>
> — Nina George, *The little Paris bookshop*

2.1 Introduction

If your friend is experiencing early risers' smugness, how would you know? From a quick glance at her face and posture, you see she is experiencing a low-arousal positive emotion. To refine this attribution, though, you would need knowledge of the context and cause of the emotion. She is more likely to feel smug, you know intuitively, if she chose to wake up early (rather than being woken involuntarily by a screaming baby) and if she used those extra hours to her relative advantage (rather than wasting them counting sheep). As this example illustrates, human observers can recognize and reason about highly-differentiated, or fine-grained, emotions. Here we propose that

fine-grained emotion concepts are best captured in a Bayesian hierarchical generative model of the intuitive theory of other minds.

The role of concepts in emotion has been extensively disputed (Adolphs, 2016; Barrett, 2014; Moors, 2014; Tracy, 2014). This question is particularly hard for *first person* emotions: when I myself feel anxious, what is the role of my concept of "anxiety" in the construction of my experience? Here, we selectively tackle an easier problem: the problem of other minds. We recognize anxiety in our friends, distinguish their anxiety from their disappointment or regret, and try to respond in appropriate ways (Zaki & Williams, 2013); but how do we make such specific and accurate emotion attributions to another person? In order to formally address that question, we situate emotion concepts in a computational model of the intuitive theory of mind (Baker et al., 2017; Lake et al., 2017).

2.2 Situating emotion concepts within an intuitive theory of mind

Initial scientific descriptions of an "intuitive theory of mind" focused on its application to predicting others' intentional actions (Wellman, 2014). Minimally, intentional actions can be predicted (and explained) as consequences of the agent's beliefs and desires, and modeled as inverse planning (Baker et al., 2009). Note that intuitive (or "lay") theories are causally structured, but generally not explicit, declarative, or introspectively accessible (Murphy & Medin, 1985). Subsequent models have considerably extended this basic premise to capture causal relations between other kinds of mental states. For example: Greg's choices additionally depend on (what he believes about) the costs of his actions (Jara-Ettinger et al., 2015); his beliefs update in response to new evidence (Baker et al., 2017); his actions are influenced by his habits (Gershman et al., 2016); and so on. A hierarchical Bayesian model of this intuitive causal theory can explain both observers' forward inferences (predicting Greg's actions given his beliefs and desires) and inverse inferences (inferring Greg's beliefs and

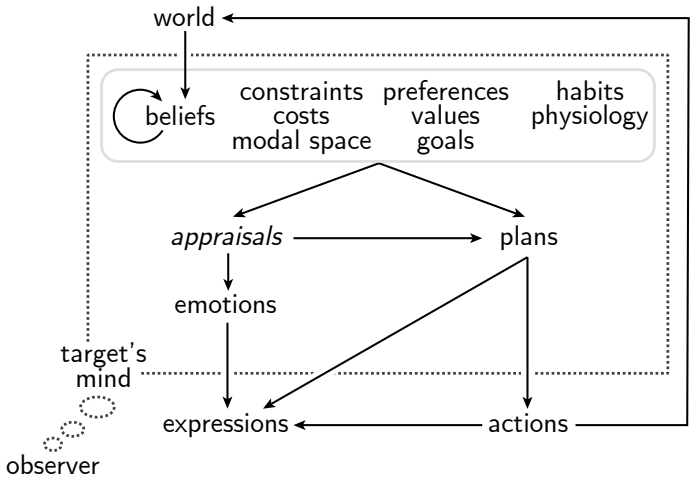

Figure 2-1: A graphical simplification of part of the intuitive causal theory of other minds. Beliefs are hierarchical and recursive (indicated by a self-referential arrow) and mediate knowledge of the world. Constraints include evaluations of the costs of actions, the modal manifold (what is possible and probable), what is controllable, etc. Preferences include both local goals and intentions, and also long-term values like morals, relationships, and status. Expressions reflect emotions, but not exclusively. Observers know that people can intentionally modulate expressions, enhancing cues of experienced emotions, suppressing expressions to mask internal states, and pragmatically effecting conformations for communicative purposes. In addition, observers know that people use their musculature for more than expressions, and readily filter for desired information using contextual knowledge about acts (he is blowing out a candle), habits (she smiles when nervous), and physiological influences (he has diminished expressions owing to Parkinson's). At the core of the model is the *inferred appraisal* process: interpreting external events through the lens of their relevance for one's goals, beliefs, costs, and so on. Inferred appraisals cause emotions (internal states) which cause expressions (observable behaviors). An observer can therefore predict emotions based on inferred appraisals (following the causal arrows) or from the observed expressions (inverse of the causal arrows). The dimension of time is implicit (variables update and inferences can be made across time). Compare similar models in Böhm and Pfister (2015), Ong et al. (2015), and Wu et al. (2018).

desires given his actions) (Baker et al., 2009).

People readily incorporate emotions in their intuitive reasoning about other minds (Ong, Zaki, & Goodman, 2016) but only recently have computational models of theory of mind been elaborated to include emotion concepts. Minimally, in the intuitive theory, emotions (or emotional reactions) are caused by how the person interprets (or "appraises") external events in relation to his other mental states (goals, beliefs, moral values, costs, traits, etc.; Figure 2-1). For example, Greg's emotional reactions will depend on whether (according to Greg) external events fulfill his goals, contradict

his beliefs, reduce the constraints or costs of his preferred actions, violate his values, and so on. As with intentional actions, the same intuitive theory also supports inverse inferences. In the intuitive theory, emotions (which are internal mental states) cause emotional expressions (which are externally observable behaviors), so observers can use perceived emotional behaviors to infer underlying emotions (i.e. perform an inverse inference from observed effects to unobserved cause). Situating emotion concepts within the intuitive theory of mind in this way may seem obvious, but has many implications for building models of human emotion understanding.

2.3 Specificity and development of emotion inference

First, this approach offers a natural, systematic way to formalize highly-differentiated predictions of others' emotions, and the links between those predictions and the rest of our sophisticated reasoning about other minds. Although no existing model has yet fulfilled this promise, parts of the intuitive theory of mind have already been well-described in Bayesian generative causal models (Jara-Ettinger et al., 2016; Moutoussis et al., 2014). Capitalizing on this progress, the same formalizations can be used to model (some) human emotion predictions. For example, in a simple lottery context, two parameters of the target's appraisal could be inferred directly from a description of the event — his overall reward, and his prediction error — and combined to capture in quantitative detail the emotions that observers predicted (Ong et al., 2015). Relatedly, Wu et al. (2018) showed participants simple moral scenarios, in which Grace puts white powder in another girl's coffee. The powder turns out to be poison, and the girl dies. Participants use Grace's smiling facial expression to infer both that Grace knew the powder was poison, and that she wanted the girl to die. These inferences could be precisely described as inverse inferences in the participants' intuitive theory of mind (Wu et al., 2018). In the real world, observers make similarly momentous inverse inferences based on emotional reactions (Armstrong et al., 2016;

"The Weeping Oscar Pistorius and a Final Question: Has It All Been an Act?", 2014).

Even children's earliest understanding of others' emotions implies (simple) inferred appraisals. Based on an agent's observed motion path (and a principle of rational action), preverbal infants can infer the agent's goal (e.g. to get over the wall); then, relative to that goal, infants can distinguish between outcomes that the agent would appraise as goal-consistent or not (Gergely & Csibra, 2003). Critically, by 10 months old, infants also appear to predict a relevant emotion (or affective state) that causes subsequent expressions (laughing or crying) and are surprised if the agent whose goal was fulfilled then shows a negative-valence behavior, crying (Repacholi et al., 2016; Skerry & Spelke, 2014). During development, children's intuitive theory of mind becomes more sophisticated, and their third-person emotion attributions follow suit (Harrison et al., 2020; Nelson et al., 2012; O'Brien et al., 2011; Ong, Asaba, & Gweon, 2016; Ornaghi & Grazzani, 2013; Ronfard & Harris, 2014; Weimer et al., 2012). Note that while some developmental psychologists reserve the term "theory of mind" for a meta-representational understanding of beliefs, e.g. O'Brien et al., here the Bayesian model of theory of mind is a generative causal theory, encompassing goals and actions as well as beliefs, costs, and values (Baker et al., 2009; Ong, Asaba, & Gweon, 2016).

The long-term goal, however, is not just to capture one or two components of observers' emotion knowledge; rather, it is to develop a formal model that captures all of the same inferred appraisals as human observers do. Promising for this line of work, when given human labels for the target's appraisals, computational models can already capture a relatively wide and differentiated range of human emotion predictions. Two recent studies provide converging evidence. Using human ratings for 25 appraisal features, a model correctly chose an emotion label (out of 14) for 51% of 6000 real-life events; only 10% of the model's choices were judged "wrong" by human observers (Scherer & Meuleman, 2013). Similarly, using human ratings for 38 inferred appraisals, a simple model correctly chose the emotion label (out of 20) for 57% of 200 short stories; human accuracy on the same test was 63% (Skerry & Saxe, 2015). These models do not yet capture the link from the event to the target's values, goals, beliefs, and costs, and thus to inferred appraisals. Still, the models' success

suggests that once these links are included, the intuitive theory of mind will capture a substantial portion of shared human knowledge about emotions.

2.4 Ambiguous perception and precise predictions

Second, our proposal offers novel insight into predictions based on combinations of inferred appraisals (forward inference) and perceived emotional expressions (inverse inference). People intuit that faces contain the most revealing information about others' emotions (Aviezer et al., 2012b; Bonnefon et al., 2013). Perhaps surprisingly, mounting scientific evidence shows that human emotion attribution from faces is actually uncertain, noisy, and low-dimensional (Bartlett et al., 2014; Hassin et al., 2013; Russell, 2016). Many different emotions can be attributed to the same facial configuration (Aviezer et al., 2015; Russell et al., 1993; Wenzler et al., 2016; Widen & Russell, 2010); and the space of emotions perceived in faces can be captured in just a handful of dimensions (Dobs et al., 2014; Mehu & Scherer, 2015). Even the valence of the event (goal-congruent or not) is not reliably perceived in high-intensity faces: the exact same facial configuration can be attributed to extreme joy (the unexpected return of a child from military service), extreme distress (witnessing a terrorist attack), extreme pleasure (orgasm), or extreme pain (unanesthetized nipple piercing) with equal plausibility (Aviezer et al., 2012b). To disambiguate these emotions, observers rely on body posture (open arms, lifted chest; Martinez et al., 2015) or inferred appraisals of the event ("he won the race"; Kayyal et al., 2015).

Although both body posture and event information are known to disambiguate emotion recognition (Aviezer et al., 2012b; Hassin et al., 2013; Kayyal et al., 2015; Zaki, 2013), our model makes a novel distinction between inverse inferences (from bodies) and forward inferences (from event-appraisals). On one hand, observers intuitively infer a common cause (an underlying emotion) of observable face, body, and vocal cues. Thus, integrating facial and body configuration, as well as vocal tone,

can improve the reliability and specificity of inverse inferences (de Gelder et al., 2015; Martinez et al., 2015; Schlegel et al., 2012). Postural information is less ambiguous than facial configuration when perceived at high intensity, from a distance, etc. (Martinez et al., 2015); similarly, vocal bursts are more informative for distinguishing among positive emotions (Simon-Thomas et al., 2009). As a result, depending on the context, the modality with the most reliable information will appear to dominate emotion attributions (Ong et al., 2015; Zaki, 2013); when one cue is ambiguous, cues from other modalities can "sharpen" the inferred cause by shifting attributions among similar, or nearby, emotions (Hassin et al., 2013). On the other hand, event information is intuitively relevant to the cause of the emotion, rather than its consequences. Additional event information can make emotion attributions more reliable not only by continuously shifting among similar emotions, but also by selecting among separated possibilities (Ma et al., 2009), because partial event knowledge can generate predictions of distinct (dissimilar, non-overlapping) alternative emotions (e.g. how will he feel after he asks his crush on a first date?). This difference between forward and inverse inferences has been obscured in prior research that confused postural and event-context cues: for example, a photograph of nipple piercing (Aviezer et al., 2012b) contains mainly event information supporting inferred appraisals, not an emotional posture.

Relatedly, we can distinguish between two ways that "dynamic" facial expressions contain more information than static ones (Krumhuber & Scherer, 2016). On the one hand, dynamic change can more precisely differentiate expressive from structural facial features (e.g. a person with dark brows from a person making an angry expression) (Goren & Todorov, 2009; Hehman et al., 2015; Todorov & Porter, 2014; Todorov et al., 2015). Dynamic change can also provide more clarity on mixed expressions, by separating the mixture in time (Jack et al., 2014). In these ways, dynamic expressions may lead to more specific or more confident inverse inferences (though observers can also be surprisingly insensitive to dynamic information *per se*; Wenzler et al., 2016; Widen and Russell, 2015). On the other hand, when temporal change in the face coincides with temporal change in the external event structure, dynamics support

forward inference by highlighting the emotionally-relevant aspect of an event (Mumenthaler & Sander, 2015). For example, observers are generally quite insensitive to elements of surprise ("wide-eyed") in mixed expressions (Mehu & Scherer, 2015; Wu et al., 2018). When a change of expression is temporally coincident with an event outcome, though, observers accurately infer that the information was unexpected and change their inferred appraisals accordingly (Wu et al., 2018). The temporal sequence of emotions can further constrain inferred appraisals; if people intuit that cognitive processes occur at different speeds then the order of expressions can indicate which hidden mental variable is associated with which emotion.

Third, we propose that there is a key asymmetry between forward and inverse inferences of emotion. The forward inference depends on inferred appraisals which are highly differentiated and granular. However, people's intuitive theory of mind is also biased and based on simplifying heuristics, inducing systematic errors (Saxe, 2005). We assume people share our desires, values, norms (Coleman, 2016). We underestimate people's ability to cope, recover, and rebound from significant events (Cooney et al., 2014; Miloyan & Suddendorf, 2015). These biases in the intuitive theory of mind translate into systematic errors in predictions of emotions. By contrast, inverse inference from emotional expressions is uncertain and low-dimensional, but also relatively accurate and unbiased. Combining both sources is therefore uniquely powerful: forward inferences from inferred appraisals can suggest highly specific, granular, differentiated predictions of another person's emotions; perception of that person's expressions can confirm or contradict these predictions, allowing for rapid correction within a reduced possibility space.

2.5 Neural representations of fine-grained emotion concepts

Finally, situating emotion concepts within the intuitive theory of mind fits well with recent neuroscientific evidence. Highly-differentiated representations of others' emo-

tions are almost exclusively found in brain regions associated with theory of mind, especially in temporo-parietal and medial frontal cortex (Ferrari et al., 2016; Sebastian et al., 2012; Skerry & Saxe, 2015). These representations are abstract and amodal, generalizing across emotions inferred from stories, events, and expressions (Skerry & Saxe, 2014, 2015). By contrast, perception of emotional expressions, and even integration of those expressions across modalities, depends on distinct brain regions, especially the superior temporal sulcus (Kotz et al., 2012; Lee & Siegle, 2014; Srinivasan et al., 2016; Watson et al., 2014; Wegrzyn et al., 2015). These two processes are dissociable in individual differences (Anderson et al., 2015; Rice & Redcay, 2015; Rice et al., 2014), and in neurodegenerative disorders (Lindquist et al., 2014). Taken together, these lines of evidence strongly support the link between emotion concepts and the rest of an observer's intuitive theory of mind.

2.6 Conclusion

Two lines of scientific research have made substantial progress in parallel, and now stand to make even more progress in concert. On the one hand, formal computational models have begun to capture the core of people's intuitive theory of mind. These models can accurately model inferences over continuous quantitative variables, within abstract hierarchical structures. As of yet, however, these models have made limited progress in the domain of emotion understanding. On the other hand, the conceptual act theory of emotion attribution identifies the powerful influence of emotion concepts on emotion attribution (though emotion concepts are usually operationalized as words, or labels; Lindquist et al., 2015; Lindquist and Gendron, 2013). Appraisal theory describes some of the content of shared knowledge about emotional events (though as a hand-picked and manually-coded list, rather than a generative causal model; Scherer and Meuleman, 2013). Using the intuitive theory of mind as a framework to formalize observers' inferences about a target's appraisals offers a powerful tool to capture, and even recreate in a computer (Gratch & Marsella, 2014), our detailed knowledge of how others feel.

Chapter 3

Reasoning about emotions, expressions, and events

> To achieve accurate knowledge of others, if such a thing were possible, we could only ever arrive at it through the slow and unsure recognition of our own initial optical inaccuracies. However, such knowledge is not possible: for, while our vision of others is being adjusted, they, who are not made of mere brute matter, are also changing; we think we have managed to see them more clearly, but they shift; and when we believe we have them fully in focus, it is merely our older images of them that we have clarified, but which are themselves already out of date.
>
> — Marcel Proust, *In Search of Lost Time*

3.1 Introduction

Imagine watching your friend, Luke, listen to a voicemail. As he listens, his facial expression remains fairly neutral, then he looks up at you with a mild smile. What just happened, and how does Luke feel about it? Intuitively, it is hard to be sure, but what makes this problem hard? One possibility is that this scenario describes a perceptual problem. In this view, often called "*emotion recognition*" or "*emotion perception*," emotions are signaled by patterns of expressive behavior (Keltner et al., 2019; Shariff & Tracy, 2011). To correctly recognize Luke's experience, observers

use perceptual pattern matching to interpret the dynamics of Luke's musculature in detail.

In this view, observers make nuanced and high dimensional inferences by observing others' richly informative expressions, because expressions can signal someone's current emotional state, intentions, and assessments of the eliciting situation (Keltner et al., 2019). Thus, with perceptual access to the relevant expression cues, observers should be able to reliably identify others' emotions (Cowen & Keltner, 2020). This "*emotion recognition*" view aligns with people's lay intuitions. For example, observers expect that intensely positive and intensely negative experiences should lead to highly distinctive facial expressions (Aviezer et al., 2012b). This view assumes (often implicitly) that observers' emotion knowledge is encoded as a network of transitive statistical associations between events, emotions, and expressions. For example, it is argued that observers can reliably map between expressions, emotions, and situations. Observers may be given a photo and asked to identify what events evoked the expression (Haidt & Keltner, 1999), or label the expression with an emotion word (Tracy & Robins, 2004). Similarly, observers may be given a description of events and asked to select which expression matches that situation (Cordaro et al., 2020). The theories and assumptions of *emotion recognition* have heavily influenced psychological study of how people understand others' emotions (for recent reviews, see Barrett et al., 2019; Keltner et al., 2019), as well as machine learning efforts to engineer artificial systems that rival human emotion understanding (B. Martinez et al., 2019; Dupré et al., 2020; Krumhuber et al., 2021; Yu & Zhang, 2015).

Here we propose a different idea, which we call "*emotion reasoning*": that observers make sense of the expressions they see, in terms of their conceptual knowledge. *Emotion reasoning* treats emotion understanding as causal inference over a hierarchical mental model, rather than as a *transitive network* of statistical associations. Recent work has proposed that lay people use richly structured mental models to reason about others' emotions (Anzellotti et al., 2021; de Melo et al., 2014; Ong et al., 2015; Saxe and Houlihan, 2017; Wu et al., 2021; for review, see Ong et al., 2019). These *"intuitive theories"* are abstract causal models that enable commonsense

reasoning (Gerstenberg & Tenenbaum, 2017; Gopnik & Wellman, 1994).

In line with this proposal, we suggest that human observers generate hypotheses about what Luke might be reacting to, based on a mental model of the causal relationships between events, emotions, and expressions. Contextual knowledge can dramatically influence which hypotheses are generated (Scherer & Meuleman, 2013; Skerry & Saxe, 2015). If the message is from an advertising company, Luke is likely to find the content dull. If the message is from a college Luke applied to, then the message likely contains momentous news, either good or bad. From hypotheses about the content of the message, observers predict how Luke would feel and what expressions he would make, in each case. Finally, observers match these predictions to their observations to get a posterior probability over the hypotheses: that is, an inference of which message, and which emotional reaction, best explain Luke's expressions. In this variety of *emotion recognition*, where observers' emotion knowledge is encoded as a common network of statistical associations, the distribution of emotions that observers predict a situation to elicit match the distribution of emotions that people appear to express in that situation.

Some varieties of *emotion recognition* grant that the emotions people are predicted to experience in a situation might differ from the emotions they appear to express. This view predicts that expressions are *inherently informative* independent of context. Observers might expect Luke to be disappointed and sad if he was rejected from his dream school, but if he was in fact happy (because he now had an opportunity to travel, for instance), his expressions would communicate his experience to observers, leading them to discount their predictions in favor of the unambiguous signals of relief and joy from his expressions. Observers' emotion predictions should only be relevant to their emotion understanding when the perceptual information is degraded or unavailable. For instance, Cowen and Keltner (2020) scraped a large corpus of images from Google and found that removing non-expression information (e.g. what activities people were engaged in) had minimal effect on which emotions were ascribed to pictured expressions. Since expressions are assumed to be *inherently informative* independent of context, *emotion recognition* research programs treat expressions as

the primary source of emotion information and have largely ignored the role of emotion predictions in how observers interpret expressions.

Other work that has investigated how observers integrate expressions and event context does not support the view that expression information dominates emotion predictions (Kayyal et al., 2015). When given simultaneous access to expressions and to event context, observers incorporated both sources of information into emotion judgments[1] (Anzellotti et al., 2021; Ong et al., 2015). Consistent with these findings, *emotion reasoning* predicts that human observers interpret expressions by reasoning about what causal explanation jointly maximizes the probability of events, emotions, and expressions. In the case of Luke, when the observer does not know where the voicemail is coming from, the smile could be interpreted any number of different ways. However, if the observer perhaps overhears the initial portion of the voicemail and realizes it's from Luke's dream school, this context constrains the potential causes of Luke's reaction—a mild smile might reveal profound disappointment.

Like *emotion recognition*, *emotion reasoning* theories that posit expressions are *inherently informative* grant that the emotions people are predicted to experience in a situation might differ from the emotions they appear to express. However, *emotion recognition* and *emotion reasoning* make different predictions about how observers integrate conflicting information. Whereas *emotion recognition* assumes that information signaled by *inherently informative* expressions will dominate emotion predictions, *emotion reasoning* assumes that the interpretation of expressions is an ill-posed problem that observers solve by constraining the space of possible meanings with predictions about which explanations are likely. In other words, the emotions observers predict someone is likely to experience shape what emotions they infer to be the cause of expressions. Thus, observers can incorporate useful contextual cues into the interpretation of expressions, which is advantageous when even perceptually rich expressions are ambiguous (Israelashvili et al., 2019), or when people do not produce

[1]One limitation is that these studies used static images of isolated facial expressions, including computer-generated faces, leaving open the possibility that if observers had access to perceptually richer expression information the contribution of predictions to emotion judgments would be reduced or obviated.

the expected expressions (Durán & Fernández-Dols, 2021).

These differing accounts of what and how expression information is used during emotion understanding lead *emotion recognition* and *emotion reasoning* to make contrasting predictions about the source and structure of errors. *Emotion recognition* frames emotion understanding as *accurate signal detection* where observers can accurately make nuanced and high dimensional inferences by observing others' richly informative expressions. The common assumption that inferences from expressions are accurate with regard to ground truth (Buck, 1994; Nakamura et al., 1990; Newen et al., 2015; Shariff & Tracy, 2011; Witkower et al., 2020) leads machine learning research to train models on human annotations and takes this as a proxy for the target's emotional experience. Similarly, in psychology, researchers often ascribe ground-truth value to canonical expressions and measure how accurately people identify the prescribed label. In this view, errors occur because judgments may be noisy, especially if the perceptual signal is degraded because an expression is partially occluded, low resolution, or changes rapidly.

Framed as a perceptual problem, observers would need to perceive the dynamics of Luke's musculature in detail in order to correctly recognize Luke's experience. The more sensitively observers detect subtle changes or brief flickers of muscle tone—crinkled eyebrows, dilated pupils, flushed cheeks, a hunched shoulder—the more accurately they can understand Luke's experience (Ekman, 1992; Matsumoto & Hwang, 2018). Because access to better perceptual information should enable better emotion recognition, observers should be more accurate if they can see bodily postures in addition to facial conformations (Aviezer et al., 2012a; Lecker et al., 2020) and the temporal dynamics of expressions rather than the static expression of a single moment captured by a photo (Ambadar et al., 2005; Goldenberg et al., 2022; Jack et al., 2014; Krumhuber et al., 2013; Sowden et al., 2021). Observers should make accurate high-confidence judgments about expressions that convey strong signals, and should make noisy low-confidence judgments about expressions that convey weak signals. This should produce judgments that are noisy, but accurate. The level of noise should be related to the ambiguity of the expression but non-specific, meaning that

the pattern of errors does not depend on the content of the signaled information.

By contrast, *emotion reasoning* is likely to exhibit peaked failure modes rather than non-specific noise. To generate hypotheses about how a person will feel in various plausible contexts, observers must rely on inferences about what a person wants and expects. Yet our intuitive theory of what other people want and expect is notoriously biased (Gopnik, 1993; Kruger & Gilovich, 1999; Saxe, 2005) and poorly reflects the properties of the world, resulting in model-based errors: systematic interactions between the content of observations and the mental model (Gopnik & Wellman, 1994). An inaccurate intuition about the emotions someone is likely experiencing might produce forward errors; for instance, expecting that Luke will be sad following a rejection from a college might lead us to seek confirmatory evidence and interpret his expressions as corroborating the hypothesized mental state, even when his expressions are ambiguous or incongruent (Anzellotti et al., 2021; Aviezer et al., 2012b; Kayyal et al., 2015). An inaccurate intuition about expression production might also produce backward errors; for instance, believing that Duchenne smiles imply subjectively positive experiences could lead us to infer that Luke was accepted when he was in fact rejected (Crivelli et al., 2015; Lei & Gratch, 2019; Sen et al., 2018).

This chapter argues for *emotion reasoning* as a theory of human emotion understanding. A central premise of this framework is that contextually-informed emotion predictions shape observers' interpretations of expressions and inferences of antecedent events. By contrast, *emotion recognition* treats emotion predictions as ancillary, or even irrelevant, to the understanding of expressions. *Emotion recognition* research programs vary in their specific assumptions, predictions, and aims, but they share the theoretical assumption that expressions are the primary source of information observers use to understand others' experiences. Because perceptual patterns of expressive behavior are the central focus of *emotion recognition* research programs, these theoretical frameworks and modeling efforts only coarsely articulate, or omit entirely, the function of emotion predictions. A general contribution of the *emotion reasoning* framework is that it enables one to quantitatively assess different hypothesized functions of emotion predictions, including what information predictions

contain, how predictions interact with expression information, and how predictions influence other social cognitive processes.

The theoretical basis is similar in spirit to other work that treats emotion understanding as inference over an intuitive theory of psychology (Barrett, 2017; de Melo et al., 2014; Freeman & Ambady, 2011; Ong et al., 2015; Wu et al., 2017). We extend this work by introducing a computational model that captures how observers reason over latent emotion representations to make abductive inferences. To our knowledge, this is the first formal model of abductive inference over emotions.

3.1.1 Spontaneous expressions in a high-stakes social dilemma

There are several requirements of a paradigm that aims to compare the *emotion recognition* and the *emotion reasoning* accounts of human emotion understanding. The expression stimuli should provide perceptually rich veridical information (as opposed to isolated static faces of posed expressions or computer-generated avatars). The design should afford an objective measure of ground truth accuracy (so that 'errors' are well defined) and the task should allow noisy judgments arising from perceptually ambiguous stimuli to be differentiated from systematically incorrect judgments.

We generated experimental stimuli designed to meet these criteria by artificially separating the perceptual information from context information in recordings of a televised British gameshow called *GoldenBalls*. Every episode of GoldenBalls culminates with two contestants playing a dramatic one-shot instantiation of the Prisoner's Dilemma (PD). Each player is given a choice to "Split" or "Steal" a jackpot (in standard Prisoner's Dilemma notation, to "`C`ooperate" or "`D`efect", respectively). If both decide to "`C`ooperate", they each receive half of the jackpot. If one player instead chooses "`D`efect", that player wins the entire jackpot and the other player who chose "`C`ooperate" leaves with nothing. If both players choose "`D`efect", both get nothing[2].

[2]Rapoport (1988) defines this payoff structure as a Weak Prisoner's Dilemma because the `CD` payoff confers the same monetary reward ($0) to player 1 as the `DD` payoff. Thus, with respect to a player's first-person financial payout, defecting is never harmful, but is only conditionally beneficial.

Players negotiate with each other in front of a live audience in an attempt to convince the other to make a choice that is financially disadvantageous (to cooperate). Each player makes a decision in private, then they simultaneously reveal their choices, all while being filmed.

The game is emotionally evocative by design. When the choices are revealed, players discover whether they have won or lost real, and often substantial, sums of money, and whether they have successfully cooperated, successfully duped, been duped by, or failed to dupe the other player. The TV cameras capture their spontaneous unscripted dynamic expressions, including changes of posture and expression of players' faces, shoulders, upper bodies, and hands.

The public nature of the gameshow is a defining characteristic of the event context in this paradigm. Players were fully aware that they were being observed by the opposing player, the show's host, the live studio audience, and a remote TV audience. Knowing they were observed, players may have muted, exaggerated, or otherwise regulated their expressions (Buck et al., 1992; Chovil, 1991; Crivelli et al., 2015; Ekman, 1993; Fridlund, 1991; Hess et al., 1995, 2005; Parkinson, 2005; Williams et al., 2021). We consider the public nature of the expressions to be a feature of our paradigm, not a bug. *Emotion recognition* and *emotion reasoning* are competing accounts of human emotion understanding in real social interactions. Much of the adaptive advantage ascribed to *emotion recognition* stems from being able to decode behaviorally relevant signals, and respond effectively, in ecologically relevant contexts of real social interactions (Shariff & Tracy, 2011; Tracy, 2014). Thus, a valid test of these competing accounts is how well they predict human understanding of spontaneous reactions in a salient social situation.

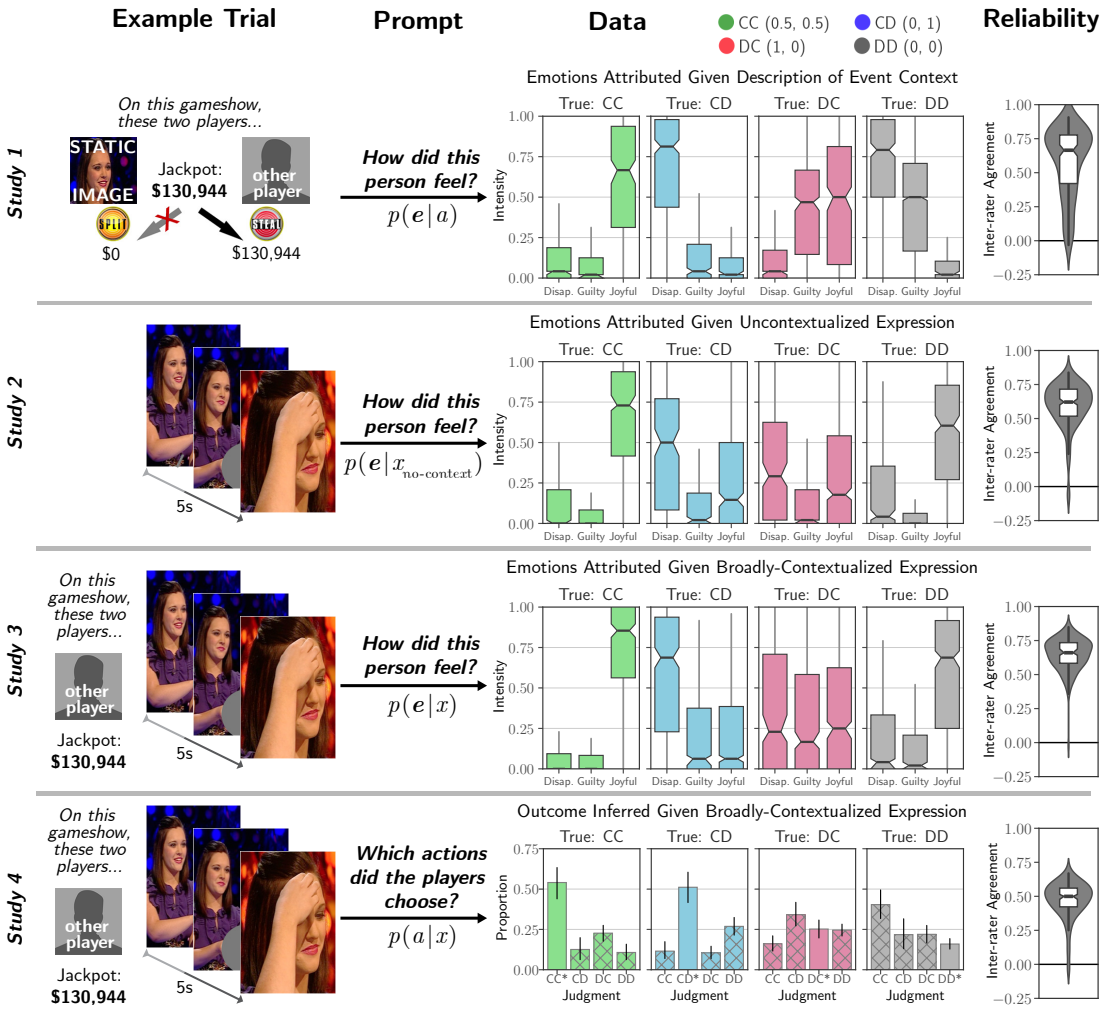

Figure 3-1: In Studies 1, 2, and 3, observers judged the intensity that players would feel 20 fine-grained emotions ($\boldsymbol{e}$) on continuous scales. Observers rated all 20 emotions during every trial. Three example emotions (*disappointment*, *guilty*, and *joyful*) are shown here (boxplot notches indicate the 95% bootstrap confidence interval of the median). In Study 1, observers predicted a player's emotions based on a description of the game outcome. The two players each decided whether to Cooperate (C), by choosing the ball with *Split* written inside, or to Defect (D) by choosing the ball with *Steal* written inside. The actions (a) of the player dyad determine how the available jackpot is paid out to the two players. In the example shown, the focal player Cooperated and her opponent Defected. Thus, the game outcome is CD and the relative payoff is $(0, 1)$ to the focal and opposing players, respectively. In Study 2, observers watched a 5-second video of a player's spontaneous expression ($x_{\text{no-context}}$) without being told anything about the broad-context of the gameshow or what events transpired. In Study 3, observers watched a player's dynamic expressions, knowing that the person was a player on a gameshow and the rules of the GoldenBalls game (x). In Study 4, observers guessed what actions the players chose (a) given the same information as in Study 3. Solid bars are correct judgments, hatched bars are incorrect judgments. Error bars give the 95% bootstrap CI of the mean proportion of judgments (n=22 videos per true outcome). Inter-rater reliability was calculated as the Pearson correlation between a single observer's judgments and the population mean across all stimuli.

3.2 Study 1: Observers make nuanced and reliable emotion predictions based on event descriptions

Emotion recognition and *emotion reasoning* ascribe different importance and functions to the role of emotion predictions in human emotion understanding. *Emotion recognition* emphasizes the importance of expressions, which are assumed to convey rich discriminative information. In most cases, the emotions someone is predicted to experience should reflect convergent information (i.e. the emotions people are expected to experience are similar to the emotions they appear to express). When predicted emotions conflict with observed expressions, then observers will discount them in favor of the rich and reliable information from expressions. Thus, the emotions someone is expected to experience in a situation will only factor into observers' emotion understanding when expression information is degraded or unavailable. *Emotion reasoning* posits that emotion predictions are central to human emotion understanding, as they constrain the ill-posed problem of inferring what mental states are likely explanations of the perceived expressions. In this first study, we began by measuring what emotions the GoldenBalls games were predicted to elicit, and if the outcomes were expected to evoke different reactions.

3.2.1 Methods

A group of English-speaking adults (N=164, 74 female) were shown descriptions of 12 out of the 88 games, then asked to predict emotions that the focal player experienced. After watching an introductory video that explained the rules of the game, observers predicted the emotions experienced by players on the gameshow. In each trial, observers were shown the size of the jackpot, the choice made by the focal player, the choice made by the opposing player, how much money each player won, and a still photo of the focal player taken before the players' choices were revealed.

Players may have chosen to Cooperate or to Defect, leading to four possible outcomes for a player dyad, which are denoted CC, CD, DC, DD. For instance, CD indicates that the focal player Cooperated, while the opposing player Defected. All the information given on a trial reflected what actually happened on the gameshow. See Figure 3-1 for a summary of the experimental design of each study. Observers predicted the focal player's emotional experience, using continuous scales to report the intensity of 20 different emotions. We hypothesized that observers would make systematic and specific predictions of what emotions players in the GoldenBalls game experienced.

3.2.2 Results and Discussion

Overall, observers expected more positive emotions when a player won money (CC and DC outcomes) and more negative emotions when a player left with no money (CD and DD outcomes). Mean emotion judgments are shown in Figure 3-2a and Figure 3-3. Observers also predicted distinct emotions based on which choice (C or D) a player made. Players were expected to feel more *guilty* and *embarrassed* when they won money by defecting (DC), but more *relieved* and *grateful* when they won money by cooperating (CC). Observers predicted players would feel *jealous* when they received no reward because they were duped by their opponents (CD), and *disappointed* and *guilty* when they received no reward because they attempted to take advantage of their opponents unsuccessfully (DD).

Independent observers tended to agree about what emotions specific players were likely to experience. We estimated inter-rater reliability by comparing the emotion predictions of one observer with the mean emotions that other observers predicted for the same players (Figure 3-1). Across emotions and players, the median Pearson correlation and 95% bootstrap confidence interval (CI) was $r = 0.66$ [0.62, 0.70], indicating that independent observers made similar predictions of players' emotion experiences.

Ratings of the individual emotions were also highly reliable across observers. When the same analysis was repeated for each emotion separately, the median correlation exceeded $r = 0.75$ for multiple emotions, indicating that independent observers

made consistent predictions about how *annoyed*, *disappointed*, and *joyful* the various players would be (Appendix Figure A-3). The emotions with the lowest reliability were *apprehensive*, *terrified*, and *surprised*, but still showed inter-rater reliability above the 95% chance level. We conducted the reliability analysis within player (across emotion) and found greater agreement on players who decided to Cooperate than on players who decided to Defect (Appendix Figure A-5). The size of the jackpot had little effect on inter-rater reliability in these data (Figure A-4).

In addition to readily differentiating outcomes, the patterns of predicted emotions are nuanced with complex dependencies. For instance, players were predicted to be *joyful* when they won money, but the amount of money players won was less important to the predictions than *h*ow the players won (by Cooperating or by Defecting). By contrast, how *disappointed* a player was predicted to feel showed little dependence on the player's choice, instead reflecting whether the player won anything and, if not, the amount of money the player could have won. Players' experiences of *fury* were predicted to depend heavily on which decisions their opponents made, but unlike *disappointment*, the intensity of *fury* was predicted to be greater for players who Cooperated than for players who Defected. Predictions of *jealousy* reflect how much an opponent won at a focal player's expense. Predictions of *guilt* depended predominantly on whether a player decided to Cooperate or to Defect and were relatively insensitive to the player's objective payoff. The only scenario in which a player was predicted to avoid *embarrassment* was by mutually Cooperating with an opponent (CC); deceiving an opponent (DC) was predicted to be embarrassing, being duped into the sucker's payoff (CD) was predicted to be even more embarrassing, and failing to deceive an opponent (DD) more embarrassing still.

While there is evidence that emotion expressions can be captured by two orthogonal bases (Kuppens et al., 2013; Russell, 1980), we find that these emotion predictions are of higher dimensionality. Even though observers generated these emotion predictions based only on event descriptions of simple games that reflect two binary choices and the size of the jackpot, four principal components are needed to capture predictions of emotions like *guilty* and *embarrassed* (see Appendix Figure A-2).

Emotions predicted by observers support classification of players' actions. We trained a linear multinomial support vector machine (SVM) to classify the game's outcome from an observer's twenty-dimensional emotion rating. The outcome results from the combination of which action (`C` or `D`) the two players in a dyad chose. When tested on held-out data, the classification accuracy achieved was 0.70 for held-out game descriptions, and 0.70 for held-out observers, both well above chance (0.25). These results indicate that the pattern of emotions predicted is highly distinct between games of different outcome categories, and that the pattern holds at the level of individual stimuli and at the level of individual observers.

Given the separability of emotion predictions based on players' actions, it is not surprising that the emotion predictions also differentiated players who won money from players who did not, and differentiated players who `C`ooperated from players who `D`efected. In left-out games, a linear SVM classified whether the focal player won money with 0.86 accuracy, and classified the focal player's action with 0.79 accuracy, where chance = 0.5.

3.2.3 Summary

Observers generated rich and reliable predictions of how players are likely to feel in different contexts. *Emotion reasoning* proposes that these predictions arise from a shared intuitive theory of other minds: a causally-structured conceptual network connecting world knowledge, theory of mind, event appraisals, and emotion concepts. Such predictions fundamentally shape how observers make sense of related perceptual information. By contrast, *emotion recognition* assumes observers preferentially rely on expressions to understand other's experiences, largely discounting emotion predictions when there is rich perceptual access to expressive signals. In Study 2, we assess what players' expressions convey to observers in the absence of context.

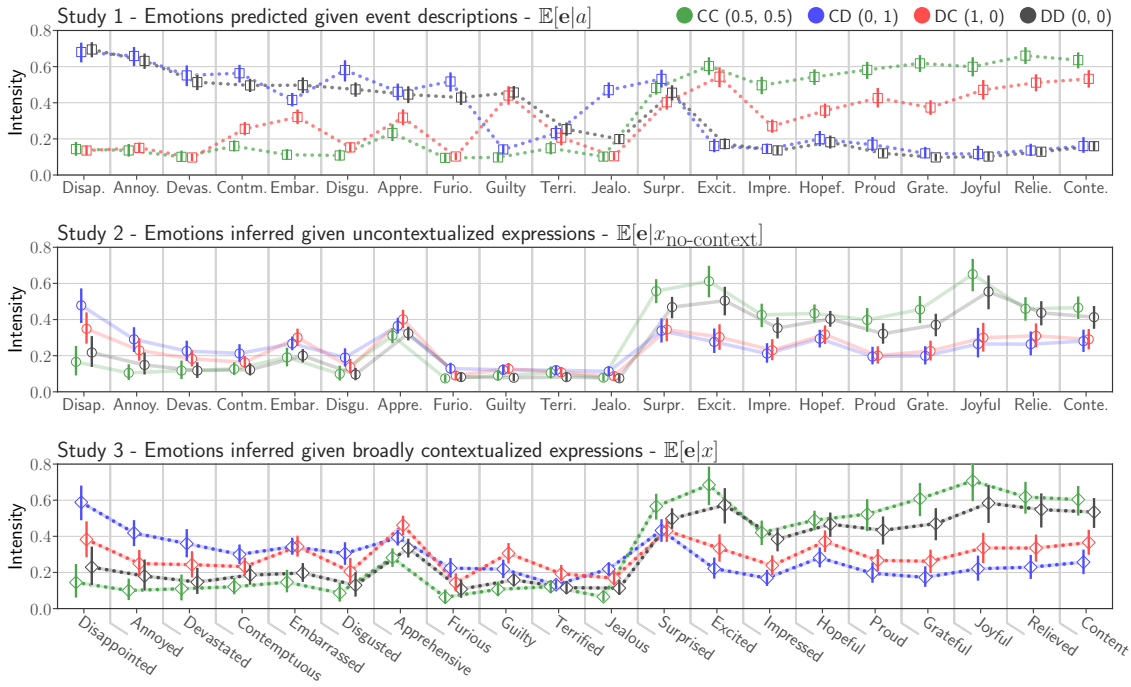

Figure 3-2: **Expected intensity judgments**. Legend indicted the colors and relative playoffs of the four outcome categories: *green* - players split the pot `CC` (focal player gets 0.5 pot, opponent gets 0.5 pot); *blue* - focal player was stolen from `CD` (focal 0, opponent 1); *red* - focal player stole from opponent `DC` (focal 1, opponent 0); *black* - both players tried to steal and got nothing `DD` (focal 0, opponent 0). Error bars indicate 95% bootstrap CI of the mean between stimuli. Connecting lines are for display purposes only, to aid in the visual grouping of related data; emotion labels are categorical so intermediate values between labels have no meaning.

3.3 Study 2: Emotions conveyed by expressions do not discriminate between event contexts

In Study 1, we established that observers expect players to experience substantially different emotions in response to the four outcome categories. In this study, we assess the patterns of emotions observers inferred when perceiving players' dynamic expressions, by measuring the emotions observers attributed to the dynamic expressions that the players produced. Under the assumption that spontaneous expressions signal players' emotions, as suggested by *emotion recognition*, observers' judgments should reflect the structure of emotional experiences evoked by the different outcomes. Therefore, players' expressions should reveal what emotions the games elicited. If the

outcomes tended to evoke different emotional reactions (as observers in Study 1 predicted), then the emotions attributed to players' expressions should be a function of the outcomes. Alternatively, the expressions might reveal that the outcomes evoked different emotional experiences in players. For instance, perhaps players' emotions were only a function of whether they won money or not, and were not a function of how they won money (cooperatively or at their opponents' expense).

By contrast, *emotion reasoning* does not assume that rich emotion information is conveyed by players' isolated expressions. Rather, *emotion reasoning* predicts observers have an intuitive theory about the generative process of expressions, which allows them to infer what mental states are likely the causes of someone's expressions. In this view, contextually-informed hypotheses constrain the inference of emotions from expressions. Thus, good hypotheses can render expression cues informative. However, without hypotheses of what emotions are contextually likely, isolated perceptual patterns of expressive behavior might be quite uninformative, ambiguous, or misleading.

3.3.1 Methods

An independent group of English-speaking adults (N=136, 55 female) saw the same players as observers in Study 1, but rather than being told what events the players experience, observers watched videos of the players' spontaneous reactions. Each video was 5-seconds and depicted the focal player's expressions using footage spliced together the moments surrounding the climactic reveal. Observers were given no information about the provenance of the videos: they were not told that the people featured in the videos were playing a televised game, the nature of the game, or the game's possible outcomes. Observers who reported familiarity with the footage were excluded from analysis. Given only the visible facial and bodily reactions, observers judged how much a player was experiencing each of the twenty different emotions.

3.3.2 Results and Discussion

In sharp relief to the emotion predictions that observers made given descriptions of the event context, the emotions that observers attributed to players given expressions were strikingly similar between the four outcome categories (Figure 3-2b). Across outcome categories, observers rated the players high on *joyful*, *excited*, *surprised* and *apprehensive*. Slightly more positive emotions were attributed to players that experienced mutual cooperation (`CC`) and mutual defection (`DD`) outcomes, and more negative emotions to players in the asymmetric outcomes (`CD`, `DC`) where a dyad made opposing decisions. The overall similarity between outcome categories is surprising in light of observers' intuitive hypotheses: observers predict that players' emotions should readily distinguish which outcomes they experienced.

As in Study 1, observers' emotion judgments were highly reliable across emotions and videos, with a median correlation of $r = 0.62$ [0.60, 0.64] between one observer's judgments and the average of other observers' judgments of the same videos and emotions (Figure 3-1). However, the pattern of inter-rater reliability differs substantially between these studies at the level of individual emotions (Appendix Figure A-3). In particular, compared to the emotions predicted by observers given descriptions of the event context (Study 1), observers of expressions showed greater inter-rater reliability in their judgments of *surprise*, but less reliability in their judgments of *furious* and *guilty*, in addition to *jealous*. The overall inter-rater reliability of emotion judgments demonstrates that observers are sensitive to perceptual cues present in individual players' dynamic expressions.

The videos that observers viewed in this study presented the dynamic face and body expressions that 88 different players spontaneously produced during monetarily and socially salient events. In comparison, the event descriptions that observers viewed in Study 1 were extremely sparse, consisting of two binary choices, the size of the jackpot, and a static photo of one player from before the choices were revealed. Despite this clear difference, the same number of orthogonal bases (four) were required to capture the emotion judgments of players' dynamic expressions, as were required

to capture the emotion judgments of the event descriptions (Appendix Figure A-2).

To confirm that the cues that observers reliably detect do not support classification of which actions players chose, we trained a linear multinomial SVM to classify the outcome of a game from the twenty-dimensional emotion attribution made by an observer. In contrast to Study 1, emotions judged from expressions support relatively poor discrimination between the four outcome categories. The classification accuracy achieved was 0.41 for held-out expression videos, and 0.41 for held-out observers (chance = 0.25).

We then tested whether a linear SVM could perform a simpler task: classifying whether or not a player won money. The classifier performed near chance, achieving an accuracy of only 0.55 (chance = 0.5). We similarly did not find evidence that emotions attributed based on players' expressions could support discrimination between players that chose to `C`ooperate and players that chose to `D`efect: the accuracy achieved for held-out players was 0.53 (chance = 0.5).

Visual inspection of the mean intensities shown in Figure 3-2 instead suggests that the structure of interpreted emotions maps onto whether or not the two players in a dyad made matching decisions. This observation is supported by the exploratory finding that a linear SVM could decode whether opposing players made matching choices (either `CC` or `DD`) versus different choices (either `CD` or `DC`), from the emotions attributed to the focal player (accuracy = 0.70, chance = 0.5).

3.3.3 Summary

Study 1 showed that predicted emotions reflect the structure of the veridical outcomes: the within-outcome similarity of emotion judgments was high whereas the between-outcome similarity was low. Given the emotions that one player was predicted to experience, the actions of both players in the dyad could be linearly decoded. In other words, observers predicted that each of the game's four outcome categories elicited distinct emotional experiences. However, in the present study, the expressions that players actually produced during the same games led observers to infer that players from different outcome categories were having highly similar emotional experiences.

The judgments of the emotions players appeared to experience were insensitive to major components of the events players actually experienced, including whether players received a monetary reward or nothing and whether players made a prosocial or a deceptive choice. Importantly, the inter-rater reliability was comparable for emotions predicted given descriptions of event context and emotions attributed given dynamic expressions (Study 1: $r = 0.66$ [0.62, 0.70]; Study 2: $r = 0.62$ [0.60, 0.64]). Taken together, these two studies show that players who were expected, from contextual information alone, to experience dissimilar emotions, instead spontaneously produced expressions that were interpreted as emotionally similar.

Finding that the emotions attributed to perceptually rich expressions differ dramatically from the emotions the players were predicted to experience does not inherently argue against *emotion recognition*. However, the pattern of judgments is not easily reconciled with the *emotion recognition* view. Although we do not know what emotions the players actually experienced, it seems unlikely that players who successfully won the entire pot (`DC` players) would, on average, be approximately as *disappointed* as their opponents who tried to split the pot and thus won nothing (`CD` players). Or that `DD` players, who won nothing, would be on average nearly as *proud, grateful, and joyful* as `CC` players, who successfully cooperated with their opponents and split the pot evenly.

3.4 Study 3: Expressions are interpreted in light of conceptual knowledge

In our view, the inference of emotions from expressions is an ill-posed problem that observers solve using a richly-structured intuitive theory. Hypotheses about what someone is likely to be feeling shape the inference of what emotional experiences are likely explanations for someone's expressive behavior. In Study 3, we test contrasting predictions about the effect of broad contextual knowledge on the interpretation of players' dynamic expressions. We repeated Study 2, this time informing observers

that the videos showed the reactions of players on the GoldenBalls gameshow. Introducing the broad context of the gameshow is a relatively subtle experimental manipulation, in that observers were still given no information about the decisions players made or the rewards they received.

Emotion recognition asserts that, with perceptual access to the relevant expression cues, observers should be able to reliably identify others' emotions independent of context (Cowen & Keltner, 2020). In other words, given players' spontaneous dynamic expressions, observers' emotion attributions should be conditionally independent of situational information. If observers' emotion attributions are conditionally independent of the broad situational cues given players' expressions, the attributions should not systematically differ from the emotion attributions in Study 2, where observers were given no situational cues. *Emotion reasoning* makes the contrasting prediction: that contextual cues constrain the space of hypotheses and shape the inference of emotions from expressions (Anzellotti et al., 2021). Thus, knowing what events are possible will induce priors over what emotions the players are likely to be expressing, and observers will make different emotion attributions than the observers in Study 2.

3.4.1 Methods

An independent group of English-speaking adults (N=135, 58 female) saw the same 5-second videos as in Study 2. Observers began the experiment by watching the same introductory video as in Study 1, which introduced the rules of the game and the possible outcome categories. Each trial provided the pot size but not the players' actions. Observers judged how much the focal player was experiencing each of the twenty different emotions. Observers who reported any familiarity with the footage were excluded from analysis.

3.4.2 Results and Discussion

Overall, the emotion attributions observers made knowing the broad-context of players' expressions were much more similar to interpretations of players' expressions

made without this context (Study 2) than to the predictions based on players' actions (Study 1). Figure 3-3 illustrates the effect of different sources of information on emotion judgments[3]. Where the patterns of judgments in this study diverge from those in Study 2, observers who knew the broad context judged players' emotions as higher intensity than observers given no context information. The effect of context information is selective–the elevation pattern is a function of the emotion label and the outcome.

In the absence of context, players' reactions to the CC and DD outcomes tended to be interpreted as low in negative emotions and high in positive emotions such as *relieved*, *joyful*, *proud*, *grateful*, and *content*. Inclusion of the broad-context magnifies these attributions, with the expressions being interpreted as even more positive. For CC videos, this tends to result in the emotion attributions shifting closer to the predicted values (collected in Study 1). For DD videos, the shift caused by context moves the emotion attributions further from the predicted values.

Some emotions do not seem to be signaled by expressions alone. In Study 2, observers were cued to *guilty* and *jealous* as possible interpretations and still reliably judged that the expressions did not convey these emotions. Knowing that the players might have lied and stolen a large sum of money (DC), or might have been deceived and fleeced (CD), led observers to interpret certain expressions as conveying experiences of *guilt* or *jealousy*.

We find that context shows complex interactions with expression information. *Furious*, for example, is not reliably attributed to players' expressions in the absence of context. When the broad-context is known, *furious* is reliably attributed to the expressions of players from CD games, but not from any other outcome. This pattern is not easily explained in terms of what was predicted and what was attributed. In

[3]In Figure 3-2, the question of interest is how representative the stimuli sampled are of the population, i.e. how the point estimate should generalize to other expressions/descriptions in a given outcome. Thus, the 95% bootstrap confidence interval (CI) for an outcome was estimated by resampling point estimates of the stimuli. In Figure 3-3, the question of interest is how broad contextual knowledge about the situation affected observers' judgments of the specific stimuli that we tested. The CI should therefore give the uncertainty in the mean of these specific stimuli, generalizing to the population of observers. Thus, the 95% bootstrap CI for an outcome was estimated by resampling observers' emotion judgments within stimulus.

both CD and DD games, *furious* was strongly predicted and not attributed, yet the broad-context led to higher attributions of *furious* for only CD expressions.

Thus, while the emotion ratings of expressions are largely similar whether or not observers are cued to what types of experiences are broadly possible, the fine-grained differences reveal complex, contextually-dependent emotion reasoning. The effect of context is not simple modulation, where context increases the attributed intensity in proportion to a baseline. The effect of context is also not the consequence of a simple interaction between the prediction and the attribution. Rather, the effect of context suggests that conceptual knowledge of the broad-context shapes the hypothesis space from which observers generate emotion predictions. As a result, observers reason about what emotions are conveyed by expression cues in light of what experiences are predicted to be likely.

As in Studies 1 and 2, observers' emotion judgments were highly reliable across emotions and videos, with a median correlation of $r = 0.66$ [0.64, 0.69] between one observer and the mean of other observers' attributions to the same players (Figure 3-1). Inclusion of the broad-context increased the reliability of most individual emotions. Notably, *guilty*, *jealous*, and *furious* were judged much more reliably when the broad-context was known (Appendix Figure A-3).

Emotion attributions made by observers that knew the broad-context of the gameshow (this study) supported only slightly better classification of the outcome than attributions made by observers who were given no context information (Study 2). The classification accuracy achieved was 0.44 for held-out expression videos, and 0.45 for held-out observers (chance = 0.25). The confusability pattern for emotions attributed to broadly contextualized expressions is also similar to those attributed in uncontextualized expressions in Study 2. Namely, emotions attributed to CC expressions are not readily discriminated from emotions attributed to DD expressions, and emotions attributed to CD expressions are not readily discriminated from emotions attributed to DC expressions. Whereas, emotions attributed to players who made symmetric decisions (CC and DD outcomes) are discriminable from players who made asymmetric decisions (CD and DC outcomes).

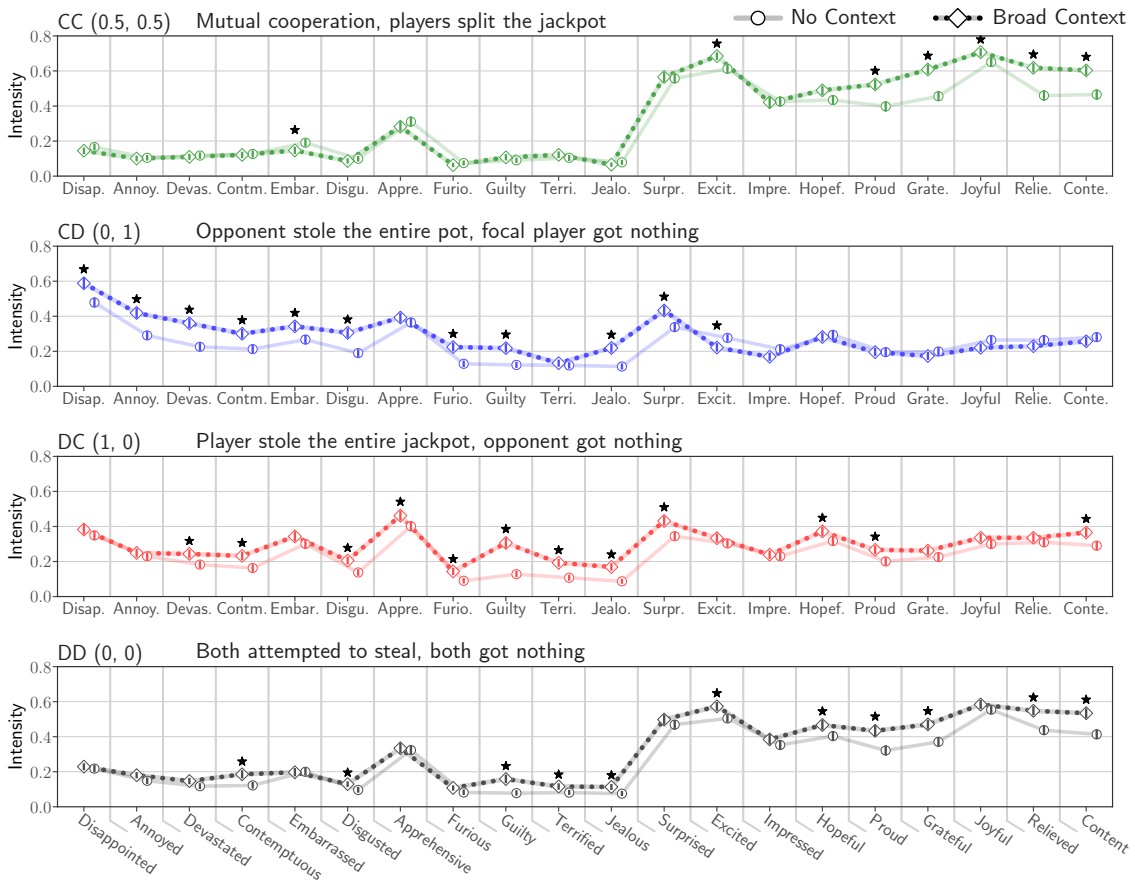

Figure 3-3: **Effect on contextual knowledge in the interpretation of perceptual information**. The point estimates in this figure are identical to Figure 3-2, but the different experiments are juxtaposed by the four outcomes. Rows show the expected judgment intensity for the four outcomes. **No Context** Shaded lines show the emotions interpreted from expressions when observers were given no information about the rules of the game (Study 2). **Broad Context** Dotted lines with shading are emotions interpreted from expressions when observers knew the rules of the game but not what specific events players experienced (Study 3). Error bars show the 95% bootstrap CI of the mean between judgments, within stimulus (note that the CI are different than Figure 3-2). Stars indicate non-overlapping CI. Connecting lines are for display purposes only.

3.4.3 Summary

Inclusion of broad contextual knowledge produced fine-grained differences in the interpretation of perceptual information. For instance, even though emotions like *guilty* and *jealous* were judged as absent from the expressions in Study 2, they were reliably attributed to players when observers could draw on conceptual knowledge to constrain the space of likely emotion experiences. While knowing the gameshow context did

bias attributions of emotions to the players' expressions, it did not appear to render the expressions more diagnostic of the events players experienced. In subsequent sections, we demonstrate the effect of broad-context on the interpretation of perceptual expression information is crucial to capturing how observers use expressions to make causal inferences.

3.5 Study 4: Systematic errors in human causal reasoning about the antecedents of expressions

The preceding studies raise the question of how observers make sense of players' expressions in a naturalistic emotion understanding task: inferring the cause of another person's emotional expressions. In the view of *emotion recognition*, observers' emotion knowledge reflects a common network of associations. Rich perceptual access to expressions should allow observers to make accurate, albeit noisy, inferences of emotions, intentions, and eliciting situations. However, Studies 1 and 2 showed a striking disconnect between the emotions observers predicted would be elicited by a situation and the emotions they interpreted from players' expressions, indicating that observers will not be able to accurately match the perceptual patterns of expressive behavior to the eliciting situations. Moreover, while observers predicted that the four game outcomes should elicit distinct emotions (Study 1), players' dynamic expression do not appear to convey discriminative information about which outcome elicited their reactions (Studies 2 and 3). This suggests that expressions alone convey less situational information than *emotion recognition* posits. It is nonetheless possible that expressions do convey diagnostic information and that our measurements simply failed to capture the discriminative information communicated to observers, leading to poor SVM outcome classification from the emotion judgments of the expressions. For instance, the diagnostic information might not be represented as emotion knowledge, or observers might not be able to report the information they decode from

expressions. It is also possible that the emotion scales we used did not query the relevant knowledge[4]. In Study 4 we specifically test if expressions convey diagnostic situational information by asking observers to directly infer game outcomes from players' expressions.

Observers were informed of the game's rules, viewed dynamic expressions, and inferred which outcome elicited each player's reaction. This paradigm mirrors Study 3, except that observers judged what situation evoked players' expressions rather than what emotions evoked players' expressions. Each expression video is veridically associated with an outcome of the *GoldenBalls* Prisoner's Dilemma. The outcomes (`CC`, `CD`, `DC`, `DD`) are determined by the decisions of the player dyad, so players' actions provide a ground-truth measure of the accuracy of observers' outcome judgments. Testing how observers recover the veridical causes of players' dynamic expressions enables a strong comparison of the behavioral predictions of *emotion recognition* and of *emotion reasoning*.

In the view of *emotion recognition*, where outcome judgments reflect perceptual pattern matching, errors should reflect perceptual ambiguity. Accordingly, incorrect judgments should statistically arise when high perceptual uncertainty forces an observer to choose between multiple competing judgments. The probability that a judgment is correct should be higher when a strong signal leads an observer to be confident in their assessment. Errors that do occur should reflect non-specific judgment noise. Judgment noise might be random and unbiased, or might show simple response bias, where the decisions have different prior probabilities (e.g. observers may be more likely to respond `CC` than `DD`). In both cases, the probability of an error depends on the strength of the perceptual signal, but the identity of the error is non-specific, meaning that it does not reflect complex interactions with the content of information signaled. Thus, judgment errors should arise from the noisy detection of ambiguous perceptual signals, but errors should not systematically arise in association with content of stimuli.

[4]This is potentially mitigated by the scope of emotion judgments we collected (continuous ratings of 20 nuanced emotions) which are far less restrictive than the most popular emotion scales.

Emotion reasoning posits that observers can additionally make systematic model-based errors. In the view of *emotion reasoning*, where outcome judgments are hypothesized to reflect model-based reasoning, certain expressions might lead independent observers to reliably and confidently endorse specific incorrect outcome judgments. Accordingly, incorrect judgments should arise when the mental model makes strong predictions that do not align with the true generative process in the world. Whereas non-specific noise (unbiased random noise and simple response biases) should be positively associated with the ambiguity of expression cues, systematic model-based errors should be inversely associated with cue ambiguity. For errors that reflect model-based misinterpretation of expressions (rather than perceptual ambiguity), observers might be highly confident in certain incorrect judgments.

Emotion recognition and *emotion reasoning* make contrasting predictions about the source and structure of outcome judgment errors. We therefore examined three sources of errors in observers' outcome judgments—random unbiased noise, simple response bias, and systematic errors—and evaluate observers' self-reported confidence in their responses. In this section (Study 4), we show that observers reliably and confidently made causal judgment errors not explained as random unbiased judgment noise or simple response biases. In the subsequent section, we will show that observers' casual judgments can be explained as model-based reasoning that produced systematic errors.

3.5.1 Methods

An independent group of English-speaking adults (N=93, 46 female) judged which outcome elicited players' dynamic expressions. As in Studies 1 and 3, observers first watched an introductory video explaining the rules of the game. Then, over 88 trials, they were shown each 5-second expression video from Study 2 along with the corresponding pot size. For every video, observers guessed which decision (`Cooperate` or `Defect`) the two players in that game had made. Observers reported confidence in the judgment of each player's action on a 3-point scale ("not confident"→ 0.0, "somewhat confident"→ 0.5, and "very confident"→ 1.0). We used the product of the

confidence ratings of the two players' actions as a summary confidence of the overall outcome judgment. Thus, if an observer reported that she was "not confident" in her judgment of one player's action, the summary confidence in her outcome judgment would be 0.0, regardless of her confidence in her judgment of the other player's action. An outcome judgment would receive the maximum confidence rating of 1.0 only when an observer reported maximum confidence in her judgments of both players' actions. Observers also estimated how *other* observers would judge the outcomes of each video.

3.5.2 Results

On average, observers performed above chance but strikingly poorly, inferring the correct outcome for an average of 36.6% [35.0, 38.1] of the videos (95% bootstrap CI estimated by resampling observers). This corresponds to a median F-score (macroaveraged across the four outcome categories) of 0.350 [0.335, 0.361], which is low but significantly greater than chance (two-sided Wilcoxon signed-ranks test, $z = 7.945$, $p < 0.001$). The population null distribution was estimated using the scores expected from shuffling judgments within observers. This approach compares each observers' performance to the performance expected if responses were generated by an observer-specific simple response bias. The mean ROC-AUC (macroaveraged across outcomes) was 0.58 [0.57, 0.59]. We estimated reliability in a fashion similar to the previous studies, by correlating each observer's judgments with the population mean (the categorical outcome judgments were expressed as one-hot encoded vectors). Across all expression videos, observers' outcome judgments showed a median correlation of 0.50 [0.48, 0.51].

Observers' classification performance was highly heterogeneous across outcomes and items. Figure 3-4 shows how well observers classified the expression videos, grouped by the outcome that players actually experienced. See Appendix Table A.1 for the numerical classification metrics and statistics. Figure 5-7 in Chapter 5 shows the item-wise outcome judgments for each expression video. When the two players had, in reality, chosen to Split the pot (`CC`), observers accurately classified 54.1% [50.6, 57.5] of videos on average (median F-score and Wilcoxon test: 0.517 [0.491,

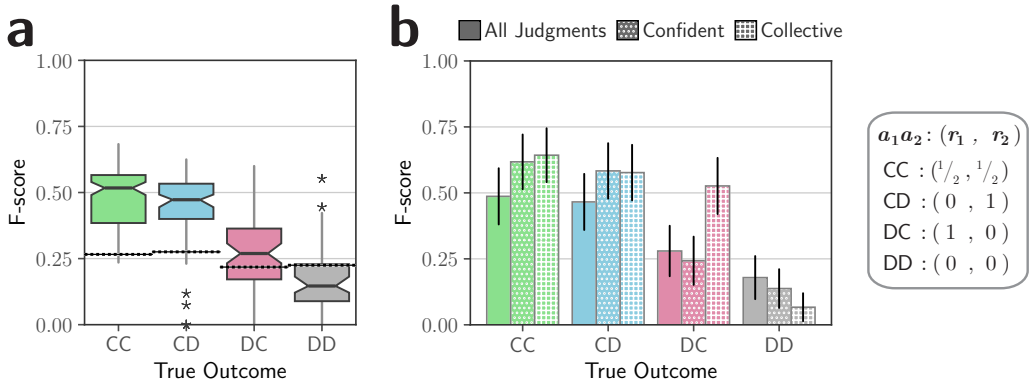

Figure 3-4: **Poor outcome recovery from expressions. (a)** F-scores of observers stratified by the true outcome. Notches indicate 95% bootstrap CI, dashed lines give the median chance performance for each outcome, based on observers' response biases. **(b)** Judgments are collapsed across observers. Error bars are binomial standard error and colors correspond to the actual outcome of the games. ***All Judgments*** gives the population F-scores, ***Confident*** gives the F-scores of the maximally confident judgments, ***Collective*** gives the F-scores of the modal judgment for each expression. The legend gives the four possible outcomes as associated monetary rewards. An outcome (a) comprises the actions chosen by player 1 and player 2 (a_1 and a_2) and the associated monetary rewards (r_1 and r_2 indicate the proportion of the jackpot paid to player 1 and player 2).

0.540], $z = 8.224$, $p < 0.001$). By contrast, when both players had, in reality, tried to Steal the pot (`DD`), on average observers accurately classified only 15.9% [13.2, 18.8] of videos, substantially *below* chance (F-score = 0.146 [0.121, 0.182], $z = -4.041$, $p < 0.001$). The classification of `DD` videos was similarly below the level of uniform random chance ($z = -5.910$, $p < 0.001$). We tested if observers' outcome judgments differentiated players who had won money from players who had not: from the focal player's dynamic expression, observers classified the opposing player's action (a_2), with accuracy = 58.4% [57.4, 59.5], and ROC-AUC = 0.58 [0.57, 0.59].

Some observers were better at this task than others. Individual differences in accuracy were stable across iterative splits of the videos (Pearson r and 95% bootstrap CI = 0.557 [0.412, 0.632]). Yet, even the best observers misclassified true events from simultaneously recorded emotional expressions almost half of the time. Of all N=93 observers who completed this task, the highest proportion of correct judgments from any participant was 52.3%.

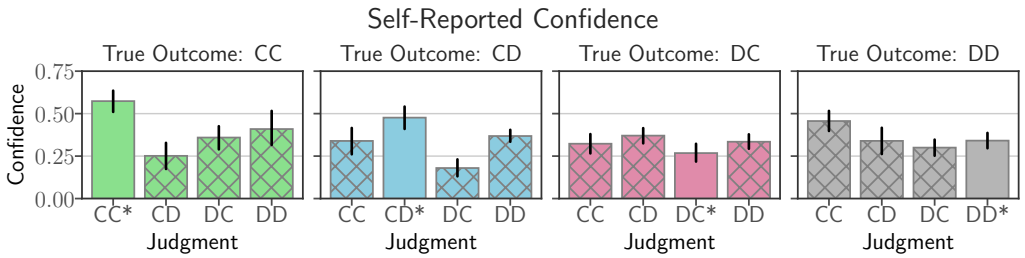

Figure 3-5: **Self-reported confidence in outcome judgments.** For each expression video, observers reported how confident they were about the two players' decisions. We calculated a summary confidence for the outcome judgment by taking the normalized product of the two confidence ratings. For `CC` and `CD` videos, observers tended to report higher confidence in their correct judgments. By contrast, for `DD` videos, observers were more confident about their incorrect `CC` judgments than in their correct `DD` judgments. This is consistent with the view that errors arise from model-based reasoning, and argues against the possibility that errors arise from `DD` videos being weakly informative. Colors indicate the true outcome associated with the expression videos and the x-axis indicates which outcomes observers inferred. Correct judgments are indicated with asterisks and are shown as solid bars. Incorrect judgments are shown as hatched bars.

3.5.3 Confidence accentuates intuitive theory

A summary of observers' confidence ratings is given in Figure 3-5. Observers tended to be more confident in their correct judgments of `CC` and `CD` expression videos, but this was not the case for `DC` and `DD` expression videos. For `DD` expressions, observers reported higher confidence in their incorrect `CC` judgments (mean and 95% bootstrap CI = 0.456 [0.398, 0.516]) than in their correct `DD` judgments (0.341 [0.296, 0.387]).

Responses in which observers were 'very confident' in their inferences of both players' actions (30.9% of all judgments) yielded better classification of the expressions of players from `CC` and `CD` games. However, confidence trended in the direction of *worse* classification of expressions from `DC` and `DD` games. Thus, when observers were maximally-confident in their assessment about what events were causally implicated by a player's expressions, they trended towards better inferences of the outcomes that tended to be correctly identified, but worse inferences of the outcomes that tended to be misidentified. Figure 3-4a shows the classification performance of the videos by outcome.

Between observers, we found no correlation between high confidence and classi-

fication performance (number of maximally confident judgments and macroaveraged F-score; Spearman $r_s = -0.013$, $p = 0.899$). Within observer, maximally confident judgments were associated with better performance when players were reacting to CC and CD games (two-sided Wilcoxon test, CC: $z = 7.127$, $p < 0.001$; CD: $z = 6.302$, $p < 0.001$). Confidence was not related to performance for DC outcomes ($z = -1.700$, $p = 0.089$). For DD outcomes, however, confidence was *inversely* related to performance ($z = -4.515$, $p < 0.001$). Observers' greater confidence in their mistakes suggests that these mistakes are not due to perceptual noise or simple response biases.

3.5.4 Collective judgments

In many domains, aggregating non-expert judgments into a population average can improve accuracy. When individual judgments are independent, noisy, and unbiased estimates of a true value, pooling judgments increases the collective accuracy (King & Cowlishaw, 2007). We tested if observers showed better collective performance than individual performance by taking the most popular (modal) outcome judgment for every video. Pooling individual outcome judgments and taking the modal judgment should reduce random unbiased judgment noise, yielding fewer errors and better classification. Collective outcome classification performance is shown in Figure 3-4b, and in Appendix A.2.7. Collective judgments resulted in better classification of DC videos, but pooling did not improve classification of DD videos at all.

3.5.5 Summary

Human observers performed poorly when asked to recover the true antecedents that elicited nonverbal expressions. A critical strength of this task is that accuracy is straightforward to measure and does not depend on experimenters' normative assumptions: it is simply how often observers inferred what actually happened from players' spontaneous emotional reactions. Observers performed poorly despite viewing stimuli that, according to the *emotion recognition* account, should be most in-

formative: spontaneous, dynamic facial and bodily expressions recorded in a real, high-stakes situation. No observer correctly classified more than 53% of the expression videos and the expressions of players from DD games were correctly classified well below chance. These results strongly oppose the *emotion recognition* account of human emotion understanding, which argues that perceiving players' expressions should allow observers to successfully infer players' experiences.

This study extends prior work measuring how accurately observers infer event antecedents from genuine spontaneous expressions. Albanie and Vedaldi (2016) had human observers view dynamic facial expressions spontaneously produced during a high-reward televised gameshow ("Deal or No Deal"). Based on a player's expression, observers judged whether the eliciting event had been financially good or bad (binary classification). The dynamic expressions were more limited than the present study, framing only the face, which can be ambiguous with regard to valence (Israelashvili et al., 2019; Wenzler et al., 2016). Albanie and Vedaldi found that, in the binary forced choice, observers achieved 62% classification accuracy on average, corresponding to an average ROC-AUC of 0.71 $[0.66, 0.76]$. Despite having visual access to face and body dynamics, observers in our present study were not more successful at inferring event antecedents. In a binary classification that resembles the previous work, observers in the present study were largely unable to differentiate the expressions of players who won money from the expressions of players who did not[5] (accuracy = 58.4% [57.4, 59.5], binary ROC-AUC = 0.58 [0.57, 0.59]). This is consistent with the idea that improving perceptual access is insufficient for improving observers' ability to interpret and use expression information. Observers' successes and failures might depend more on whether or not cues from the context induce useful priors, than on better perceptual access *per se*.

Prior work used a binary choice between events of opposing valance, where the first-person financial reward defines the event structure (Albanie & Vedaldi, 2016), which makes detecting systematic classification errors difficult. The present study's

[5]In Study 2, emotions attributed these same expressions supported the SVM classification of players who won money (CC, DC) from players who did not (CD, DD) with comparable but somewhat lower accuracy of 0.55

relatively richer event structure enabled deeper insight into the cognitive processes involved in reasoning about the antecedents of expressions. A key characteristic of the current experimental paradigm is that it affords analysis of a complex pattern of causal inferences (a 4×4 matrix of *inferred* $\times$ *veridical* outcomes) where errors are objectively defined. The event categories are distinguished not only by valence, but by the interpersonal relations. The event structure evokes complex theory of mind considerations by combining selfish monetary rewards, social utility, prosocial motivations, intentional decisions of self and other, and deception. Since the four outcome categories differed along these dimensions, they were associated with strikingly different response profiles, evident in the between-outcome heterogeneity of performance, confidence, and patterns of errors.

We leveraged the event structure to test if observers' outcome judgments resemble perceptual pattern matching or model-based reasoning. Pooling independent unbiased judgments should reduce judgment noise and improve classification performance. We found that pooling judgments improved the classification of `DC` expressions but did not improve the classification of `DD` expressions. This indicates that random unbiased judgment noise alone cannot account for observers' causal judgment errors (see *Collective* in Figure 3-4b and Appendix A.2.7). Simple response biases (the tendency to make certain judgments over others) should shift judgments towards the prior probability of a judgment in inverse proportion to the informativeness of an expression, but should not produce complex logical reinterpretations of the content. We found that simple response biases do not fully account for the pattern of errors in observers' outcome judgments. The F-scores indicate that classification of `DD` expressions was significantly below the level expected when individual simple response biases were taken into account (Figure 3-4a). Additionally, if the low classification performance of `DD` videos resulted from the expressions being perceptually ambiguous, observers should be expected to report less confidence in their judgments. We found the opposite trend: for `DD` expression videos, observers reported greater confidence when they incorrectly judged that the videos were from `CC` games than when they judged the videos correctly (Figure 3-5).

The observed pattern of errors with respect to ground-truth is inconsistent with the predictions of *emotion recognition* and the idea that observers match patterns of expressive behavior to eliciting events using a single system of statistical associations between expressions, emotions, and situational antecedents. The pattern of errors is not explained by random unbiased noise and simple response biases alone, and observers reliably and confidently misinterpreted players' spontaneous expressions. In the next section (Study 5), we show that the pattern of correct and incorrect judgments can be explained as model-based reasoning.

3.6 Study 5: Abductive inference over latent emotion representations

The preceding sections (Studies 1-4) provide strong evidence against the *emotion recognition* account of emotion understanding, illustrating that even perceptually rich expressions are surprisingly uninformative in isolation. In this section (Study 5), we make a positive case for *emotion reasoning*, arguing that observers make sense of others' emotional reactions by reasoning over latent emotion representations to infer what hypothetical events provide the best causal explanation for the expressions observed. In the view of *emotion reasoning*, perceptual information (players' dynamic spontaneous expressions) and conceptual knowledge (hypotheses about what emotions are likely) mutually constrain which explanations are probable. In this way, emotions act as latent causal links between event and expressions. We formalize the guided search for a causal explanation as a "Bayesian belief updating" model. Observers use expressions to adjudicate between alternative causal hypotheses by comparing the emotions a player appeared to express against the emotions hypothetical events were predicted to elicit.

This model reflects the intuitive theory that emotions are reactions to antecedent events, and that expressions are caused by emotions. This lay causal theory can be written as a directed graph $a \rightarrow e \rightarrow x$. Outcome a is defined by the tuple

$\langle a_1, a_2 \rangle$, where a_1 and a_2 are the actions of the focal player and the opposing player, respectively, such that $a \in \{\texttt{CC}, \texttt{CD}, \texttt{DC}, \texttt{DD}\}$. Emotions e is the vector of intensities for the 20 emotions we collected. Expression x is the dynamic expression of the focal player. In this scenario (Figure 3-6), observers make an inference to the best explanation: observers causally reason over a hierarchical mental model to make an abductive inference of which outcome elicited a player's nonverbal expression.

To build a Bayesian model of observers' intuitive reasoning, we used the data from the preceding studies (see Figure 3-1). Using the Study 1 data, we formed a distribution over the emotions players' were predicted to experience given the game outcome $p(\boldsymbol{e} \mid a)$. From the data of Study 2, we formed a distribution over emotions attributed to players' dynamic expressions when no context was provided $p(\boldsymbol{e} \mid x_{\text{no-context}})$. From the data of Study 3, we formed a distribution over emotions attributed to players' dynamic expressions when observers knew the broad-context of the GoldenBalls gameshow, but not what outcome occurred, $p(\boldsymbol{e} \mid x)$. Last, we used Study 4 to create a distribution over the outcomes that observers inferred based on players' dynamic expressions $p(a \mid x)$.

3.6.1 Abductive inference model

We simulated observers' outcome predictions as inference in an intuitive theory of mind, where observers use the emotions they infer from players' dynamic expressions $p(\boldsymbol{e}|x)$, and the emotions they expect players to experience in the possible outcomes $p(\boldsymbol{e}|a)$, to infer which outcome was most likely to have generated the observed expressions (Figure 3-6a). This corresponds to inferring the posterior probability $p(a|x)$: the probability that observers infer that a player's dynamic expression x was elicited by the event a. Given a player's expression, we assume that observers make a joint inference of the player's emotions and which outcome the player was reacting to[6], $p(a, \boldsymbol{e}|x)$.

[6]This formalization makes the simplifying assumption that the outcome is conditionally independent of the pot size given a player's expression. We therefore treat $p(a|x)$ as an approximation of the posterior probability $p(a|x, pot)$. This assumption is strictly a modeling convenience, not a hypothesis about observers' intuitive theory. How observers reason about an expression likely depends on their knowledge of how much money was at stake, but for this model we assume that the variance between outcomes is more important the effect of the pot size.

The probability of observers guessing outcome a can be computed by marginalizing the joint distribution over emotions. In the graphical model of observers' intuitive theory (Figure 3-6a), actions and expressions are conditionally independent given emotions, such that $p(a, \boldsymbol{e}|x) = p(a|\boldsymbol{e})p(\boldsymbol{e}|x)$.

$$p(a|x) \;=\; \int_{\boldsymbol{e}} p(a, \boldsymbol{e}|x)\,\mathrm{d}\boldsymbol{e} = \int_{\boldsymbol{e}} p(a|\boldsymbol{e})p(\boldsymbol{e}|x)\,\mathrm{d}\boldsymbol{e} \;=\; \mathop{\mathrm{E}}_{\boldsymbol{e}\sim p(\boldsymbol{e}|x)} p(a|\boldsymbol{e}) \tag{3.1}$$

We model observers' mental distribution of emotions given expressions, $p(\boldsymbol{e}|x)$, as the empirical distribution of emotion judgments $\hat{\boldsymbol{e}}_i$. To approximate the expectation in equation (3.1), we sum over the responses from the N_x participants (indexed by i) who attributed emotions to expression x in Study 3.

$$p(a|x) = \frac{1}{N_x}\sum_{i}^{N_x} p(a|\hat{\boldsymbol{e}}_i), \qquad \text{where } p(a|\hat{\boldsymbol{e}}_i) = \frac{p(a)\,p(\hat{\boldsymbol{e}}_i|a)}{\sum_a p(a)\,p(\hat{\boldsymbol{e}}_i|a)} \tag{3.2}$$

We model observers' mental distribution of emotions elicited by players' actions, $p(\boldsymbol{e}|a)$, based on the emotion predictions $\hat{\boldsymbol{e}}_j$ from game descriptions (comprising an outcome, pot size, and player photo). Responses from the N_a participants (indexed by j) who were shown game descriptions c with outcome a in Study 1, were used to construct a weighted Kernel Density Estimate (KDE) of the emotions observers expect. We weight each response by $w_j = (V_a N_{c_j})^{-1}$ to account for the number of observers who saw each description (N_{c_j}), and the number of videos V_a with outcome a.

$$p(\boldsymbol{e}|a) = \sum_{j}^{N_a} w_j \mathcal{N}(\boldsymbol{e};\; \mu = \hat{\boldsymbol{e}}_j, \boldsymbol{\sigma} I) \tag{3.3}$$

The KDE estimates the population's marginal distribution $p(\boldsymbol{e}|a)$ (the emotions players were predicted to experience) as a weighted mixture of Gaussian kernels, where the vector $\boldsymbol{\sigma}$ is the kernel bandwidth corresponding to each emotion and I is the 20-dimensional identity matrix. The kernel bandwidth was calculated for each emotion based on the sample standard deviation using Scott's Rule (Scott, 1992).

Observers' prior over the actions chosen by a player dyad is given by $p(a)$. Ob-

servers in Study 4 were asked to estimate how the population would judge the expressions, independent of their own judgments. We use observers' guesses of the population's judgments to estimate the population's action prior $p(a)$.

The critical juncture between the emotions observers infer given expressions, and the emotions observers predict given event context, is given by the term $p(\hat{\boldsymbol{e}}_i|a)$. This is the probability that $\hat{\boldsymbol{e}}_i$, the emotion vector attributed to a player's expression, was sampled from the conditional distribution $p(\boldsymbol{e}|a)$, the emotions observers expected players to experience when the outcome was a. The resulting posterior gives the distribution of responses over outcome-categories, i.e. $p(a|x)$ is the simulated probability of observers inferring that expression x was a reaction to outcome a.

3.6.2 Results

Figure 3-6b shows the outcome judgments of human observers, and the outcome judgments simulated by the Bayesian *emotion reasoning* model. Comparing the human and *emotion reasoning* model judgments reveals a close match across all categories. Videos belonging to the four outcomes elicited different classification patterns from observers. Observers tended to infer the correct outcome from players' expressions in `CC` games, where a player won half the jackpot by mutually cooperating with the opposing player. When `CC` videos were misclassified, observers tended to choose the other outcome that would confer a financial reward (`DC`), more often than the outcomes in which the player would receive nothing (`CD` or `DD`). Observers also tended to classify `CD` videos correctly. The most common error was to misclassify `CD` video as the financially similar `DD` outcome (the player achieves the minimum payoff in both outcomes). In contrast with videos from `CC` and `CD` games, people did not tend to correctly classify `DC` videos. Observers inferred that `DC` videos showed players in `CD` games more often than the true outcome.

Observers overwhelmingly inferred that the expressions made by players in `DD` games (where the two players defect on each other and leave with nothing) were produced by players experiencing `CC` outcomes (where the two players mutually cooperate and share the jackpot). Interestingly, for `DD` videos, the correct outcome was the least

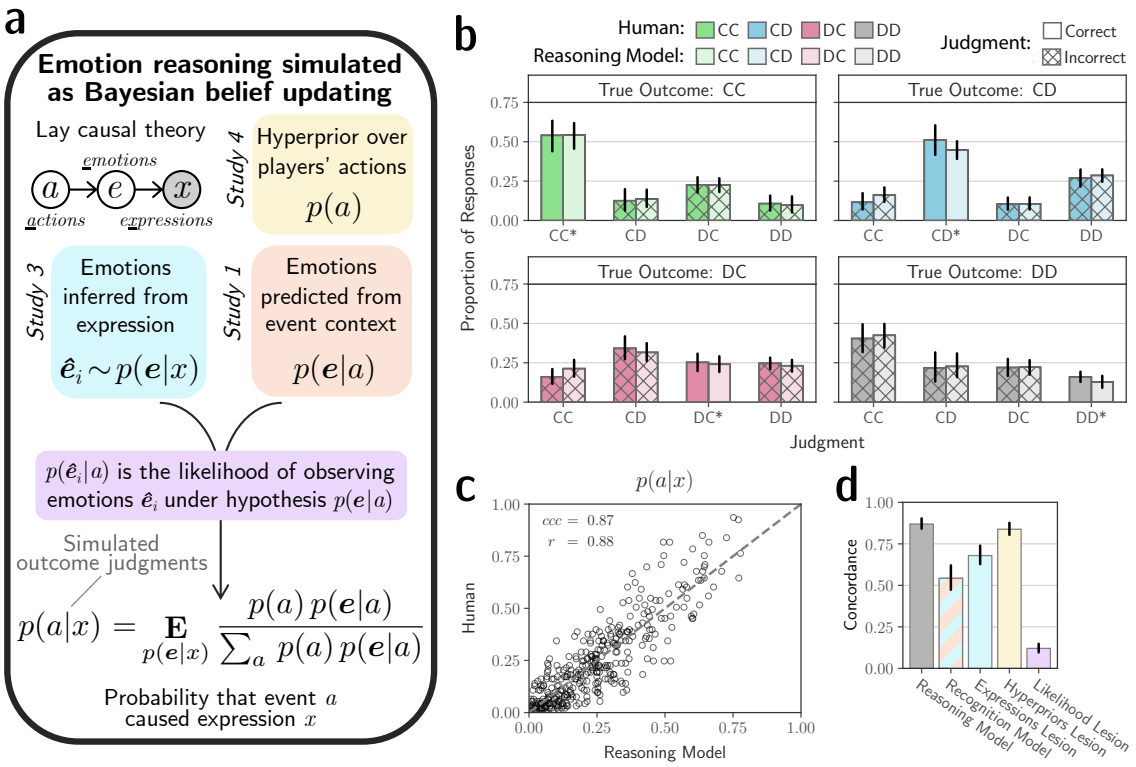

Figure 3-6: **Bayesian model of human causal reasoning**. **(a)** Abductive inference of outcomes from expressions. **(b)** Judgments of outcome grouped by ground-truth. Correct judgments are indicated by asterisks in the label and are depicted with solid bars. E.g. Observers incorrectly guessed that the expressions of players from `DD` games (bottom right cell) were from `CC` games more than the correct outcome. Error bars give 95% bootstrap CI. **(c)** Each point shows $p(a|x)$ for a given outcome a and expression x. E.g. a single point shows the model estimate of how often a given expression video would be judged to be `CC` versus the proportion of human observers who judged the video to be `CC`. **(d)** The concordance and 95% bootstrap CI of the models. Bar color indicates which part of the model in subfigure **a** was lesioned. The intact *reasoning* model is gray.

popular judgment. It may seem surprising that observers systematically misclassify `DD` videos as `CC`, but this is precisely the pattern predicted by the model.

The model similarly predicted observers' overall accuracy with respect to ground truth. Human observers judged the correct outcome with an accuracy and 95% bootstrap CI (estimated by resampling stimuli) of 36.6% [32.7, 40.3]. The *reasoning* model predicted that observers would judge outcomes with an accuracy of 33.9% [31.0, 36.8].

The aggregate data in Figure 3-6b is detailed at the stimulus-level in Figure 3-6c. Each point shows how much the *reasoning* model versus humans judged that the

specific expression video x showed a player reacting to the specific outcome a, e.g. the proportion of `DD` judgments that the *reasoning* model predicted versus the proportion of `DD` judgments human observers made, for a given video (regardless of ground truth outcome associated with the video). The *reasoning* model accurately captured the empirical data (concordance correlation coefficient $ccc = 0.869$ [0.841, 0.904]; Pearson $r = 0.876$ [0.849, 0.909]; 95% bootstrap CI were estimated by resampling stimuli with replacement across outcomes). Lin's Concordance Correlation Coefficient penalizes deviations from the identity line (perfect prediction), making it a more stringent metric of explanatory power than Pearson's correlation (Lin, 1989).

3.6.3 Lesion models

The *emotion reasoning* model simulates human causal judgments as Bayesian inference over three empirically-derived distributions: $p(\boldsymbol{e}|a)$ is the emotions players were predicted to experience based on descriptions of events (Study 1); $p(\boldsymbol{e}|x)$ is the emotions attributed to players based on their dynamic expressions (Study 3); and $p(a)$ is the hyperprior over which actions players' chose (Study 4). To test how important these components were for capturing observers' outcome inferences, we selectively "lesion" the model by replacing each one with a less structured distribution (Figure 3-6d).

The first lesion model, the recognition model, simulates the outcome judgments under the assumption that observers have an accurate understanding of what emotions triggered player's expressions. We replace the emotions observers predicted that players would experience, $p(\boldsymbol{e}|a)$, with the emotions players appeared to experience in those events $p(\boldsymbol{e}|x_{\text{no-context}})$. In the expressions lesion, we replace $p(\boldsymbol{e}|x)$, which was attributed when observers knew that the dynamic expressions x were made by players on the GoldenBalls gameshow, with $p(\boldsymbol{e}|x_{\text{no-context}})$, which was attributed to the same dynamic expressions in the absence of context. In the hyperpriors lesion, we replace $p(a)$, the distribution over outcomes, with a uniform distribution which assumes that observers considered the four events equally likely. In the likelihood lesion, we replace $p(\boldsymbol{e}|a)$ with an uninformative distribution where the predicted emotions do not de-

pend on the event, a. This represents the assumption that observers did not expect the events to reliably elicit different emotions. Whereas the full model is parameter-free, the lesion models are fit directly to the test data in order to estimate the upper extent of how well each lesion can capture human behavior (see Methods 3.8).

***Recognition* model.** A common (but often not explicit) assumption is that emotion knowledge reflects a *transitive network*, where a shared embedding of statistical associations allows observers to map between expressions, emotions, and events. The *recognition* model simulates outcome judgments under the assumption that observers have an accurate understanding of expression production. For each outcome, a, we statistically match the predicted emotion distribution to the emotions that observers judged from the perceptual information conveyed by players' expressions during that event.

$$p(\boldsymbol{e}|a) \;=\; \mathop{\mathrm{E}}_{x\sim p(x|a)} p(\boldsymbol{e}|x_{\text{no-context}}) \tag{3.4}$$

The KDE is formed in the same fashion as equation (3.3). Responses were weighted by the number of observers who saw each video and the number of videos in each outcome. Outcome judgment simulated by the *recognition model* showed a concordance and 95% bootstrap CI of 0.543 [0.474, 0.621]. The low fit to human behavior argues against framing emotion understanding as embedded knowledge of accurate associations between expressions, emotions, and events that allow observers to perceptually match patterns of expressive behavior with eliciting events.

Expressions lesion. Another assumption common in *emotion recognition* theories is that expressions are *inherently informative*. In this view, observers can reliably detect signals conveyed by patterns of expressive behavior, independent of context. By contrast, *emotion reasoning* assumes that the interpretation of expressions is an ill-posed inverse problem that observers solve by using hypotheses to constrain inference over an intuitive theory. In Study 3, we showed that broad contextual knowledge

shaped how observers interpreted players' expressions. However, those results do not address whether the effect of context is important for modeling how observers *use* expressions during social cognition.

We now test if it is appropriate to assume that the inference of emotions from expressions is independent of context. We replace the broadly contextualized emotion attributions from Study 3, $p(\boldsymbol{e}|x)$, with the emotions attributed based exclusively on the perceptual content of players' expressions from Study 2, $p(\boldsymbol{e}|x_{\text{no-context}})$. Thus, the posterior of the lesion model is given by:

$$p(a|x) \;=\; \mathop{\mathrm{E}}_{\boldsymbol{e}\sim p(\boldsymbol{e}|x_{\text{no-context}})} p(a|\boldsymbol{e}) \tag{3.5}$$

Without conceptual knowledge about the broad context of the expressions, the model fit is reduced to 0.679 [0.628, 0.739], demonstrating the necessity of capturing how even broad contextual knowledge shapes the emotions people see in expressions. The judgments simulated by the lesion model appear consistent with the empirical changes observed from Study 2 to Study 3. For example, a lack of broad contextual knowledge leads `CC` expressions to be interpreted as less *excited, proud, joyful, and relieved*, and the lesion model correspondingly underestimated how often observers correctly judge these expressions as `CC`.

Hyperprior lesion. Human observers show a statistical tendency to predict that the focal player cooperated whereas the opposing player did not (`CC` = 0.30, `CD` = 0.31 vs. `DC` = 0.20, `DD` = 0.19). To test the extent to which simple response biases are a factor in observers' reasoning, we replace the empirical prior over outcomes $p(a)$ with a uniform prior. Under the uniform action prior for both players in a dyad, the model shows a reduced match with human reasoning, dropping to $ccc = 0.838$ [0.804, 0.877]. This is evidence that, while contextually-independent response bias does factor into observers' decisions, it contributes less to the explained variance, relative to the other components.

Likelihood lesion. The core computational inference of the model compares the emotions that a player was predicted to experience during hypothetical events, with the emotions that the player appeared to express. Specifically, if an observer attributed $\hat{\boldsymbol{e}}_i$ to expression x, the model infers the likelihood that the emotion vector was sampled from $p(\boldsymbol{e}|a)$ for each outcome a. As a baseline, we completely lesion the reasoning about how emotions depend on events $p(\boldsymbol{e}|a)$, making the comparison between expressions and predictions uninformative. In this lesion model, the predicted emotions do not depend on a so any emotion vector has the same likelihood in every event. Thus, only the simple response bias $p(a)$ (the hyperprior) factors into the simulated outcome judgments. As expected, lesioning all reasoning greatly reduces how well the model matches human judgments ($ccc = 0.120$ [0.096, 0.148], 95% bootstrap CI estimated by resampling stimuli with replacement across outcomes).

3.6.4 Summary

The *reasoning* model accurately predicted that observers would systematically misclassify which events players were reacting to, and also accurately predicted how the expressions would be judged. Put simply, the model was right to be wrong, and was wrong in the right ways. The emotions that players were expected to experience, and the emotions they appeared to express, were sufficient to capture key patterns of how human observers causally reasoned about unseen events. The high agreement between the predictions of the *reasoning* model and human behavior supports our proposal that people use a flexible and sophisticated intuitive theory of emotions to reason about other minds.

Importantly, no reference to emotion was made in Study 4 (the judgments to be matched); observers guessed what outcomes players experienced and were not cued to think about emotion. Yet, by treating the emotion judgments of independent observers as the latent emotion representations of observers in Study 4, the *reasoning* model successfully predicted their outcome judgments. Thus, the 20-dimensional emotion judgments we collected contain richly structured social information sufficient to capture participants' reasoning about the causal connection between unobserved

events and others' emotional expressions.

3.7 General Discussion

We used four behavioral studies and a formal model to test the contrasting predictions of two accounts of human emotion understanding. In the view of *emotion recognition*, observers understand others' emotions by matching patterns of perceived facial and bodily configurations and gestures (Ekman, 1993; Izard, 1994). In the view of *emotion reasoning*, observers use a causally-structured intuitive theory to predict what emotions are likely in order to interpret the contextualized meanings of observed expressions (see Chapters 2 and 3); information from different sources are combined to infer a posterior probability over hypothesized emotions (Anzellotti et al., 2021; Ong et al., 2015). Our results support the *emotion reasoning* account.

Contrary to the predictions of *emotion recognition*, we find that in isolation people's spontaneous emotional expressions are quite non-diagnostic about the situation that evoked them. Observers in Study 2 attributed remarkably similar emotions to the expressions of players in situations that observers in Study 1 expected to evoke highly dissimilar emotional experiences. Similarly, in Study 4, observers were strikingly poor at using spontaneous expressions to recover the ground truth of situations. Despite being shown stimuli designed to provide the maximum amount of visually perceptible expression information, observers identified the correct events only slightly better than random guesses. Moreover, the poor group performance was not driven by a low-scoring sub-population. Rather, we found no evidence for high performance; the highest accuracy achieved by *any* observer in this 4-way classification task was 52.3%. Also contrary to *emotion recognition*, we found that broad contextual cues influenced how observers interpreted expressions in Study 3. Conceptual knowledge about what events were possible had consequential effects on how expressions were interpreted: in Study 5, ignoring the effects of broad conceptual knowledge led to a marked reduction in how well observers' outcome judgments could be explained by a Bayesian model that simulates human causal reasoning.

These results fit with a growing body of evidence that shows human observers are surprisingly bad at decoding expressions spontaneously produced during known, real-life events. To human observers, the facial expressions that people spontaneously produce during highly desirable experiences, such as winning a point in an important professional tennis match, receiving an extravagant prize, or the surprise homecoming of a deployed family member, are indistinguishable from the expressions produced during highly aversive experiences, such as losing an important point, attending a funeral, or being a bystander to a terrorist attack (Israelashvili et al., 2019; Wenzler et al., 2016). Although people intuitively expect that intensely positive and intensely negative experiences should lead to highly distinctive facial expressions, spontaneous expressions produced during intense, real-life events convey no diagnostic information about valence to observers (Aviezer et al., 2012b). This unintuitive result is highly replicable (Camerer et al., 2018). Consistent with this work, we found that observers do not effectively differentiate expressions from events of opposing valence. In Studies 2 and 4, we found high confusability of expressions from `CC` games (where players mutually cooperate and split the pot) with expressions from `DD` games (where both players defect and both lose the pot). We similarly found high confusability of expressions from `CD` games (where the opposing player betrays the focal player, stealing the entire pot) with expressions from `DC` games (where the focal player wins the entire pot).

Why are spontaneous emotional expressions, occurring in real-world, high-stakes social interactions, not diagnostic? One explanation of the prior findings might be that the relevant information is not conveyed by facial configurations, but rather by bodily postures and dynamic gestures. For instance, while observers find static photos of facial expressions ambiguous, diagnostic valance information might be signaled by bodily postures (Aviezer et al., 2012a, 2012b; Witkower & Tracy, 2019) or by the temporal dynamics of facial expressions (Ambadar et al., 2005; Goldenberg et al., 2022; Jack et al., 2014; Krumhuber & Scherer, 2016; Krumhuber et al., 2013; Sowden et al., 2021). However, our stimuli included upper bodies, arms, and hands, and contained dynamic transitions from before to after the emotion-evoking revelation.

Yet, in Study 4, observers were still remarkably poor at distinguishing even whether the evoking context was positive (winning money) or negative (losing money). Using dynamic videos of facial expressions from a different gameshow, "Deal or No Deal", Albanie and Vedaldi (2016) similarly found that observers were only 62% accurate at decoding whether players had won or lost money. Thus, human observers do not effectively use spontaneous expressions to recover the eliciting event, whether the stimuli are perceptually impoverished (face-only, static) or enriched (bodies and faces, dynamic).

A second explanation might be that spontaneous expressions are muted or mixed, leading observers to make noisy, unreliable, or uncertain inferences about emotions from these stimuli. Substantial prior evidence rules out this interpretation. Cowen and Keltner (2020) collected a corpus of 1,500 images of people by querying Google for terms like "contemptuous teenager", "yuck", and "hot tub". Online raters selected which single emotion label, from 28 fixed choices, was the best match for facial expression isolated from these images. Using split-half canonical correlations analysis, the facial expressions were reliably perceived on 27 significant dimensions (the maximum dimensionality possible). That is, observers agreed to an impressive extent on which verbal labels corresponded with which images. Cordaro et al. (2020) found that, when presented with descriptions of event antecedents, observers spanning 7 industrialized cultures reliably selected the posed prototypical expression in an 18-way unforced-choice. This body of work offers ample evidence that isolated facial expressions lead to fine-grained and reliable emotion judgments under conducive conditions. Similarly, in our data, observers showed high inter-rater reliability in emotion attributions (Studies 2 and 3) and in outcome judgments (Study 4), when given dynamic expressions. Thus, we argue that spontaneous emotional expressions can be judged reliably by observers, but that knowing how observers judge the isolated perceptual information of an expression does not explain how observers interpret and use that information for naturalistic social cognition.

Instead, we argue that spontaneous emotional expressions are not inherently informative, but can be conditionally informative with respect to specific hypotheses.

In our view, human emotion understanding should be thought of as causal reasoning in an intuitive theory of other minds. When seeking to understand another person's emotions, human observers generate predictions of what emotions are likely under different hypotheses about the situation. In our data, observers make specific predictions about what emotions people are likely to experience in each event of the game (Study 1). Prior work has similarly shown that observers readily predict the emotional reactions of people in different situations (Ong et al., 2015; Skerry & Saxe, 2015). Even when the target is not visible, much of the variance in the emotions attributed to characters in films can be predicted based on the surrounding scene (Z. Chen & Whitney, 2019, 2020).

Once a prediction is made, observers then interpret expressions they perceive in light of their hypotheses. A wide variety of contextual sources have been shown to affect interpretations of facial expressions (Atias & Aviezer, 2021; Barrett et al., 2019; Hassin et al., 2013; Hess & Hareli, 2015; Russell, 2016). When events and expressions are jointly available, emotion inferences depend heavily on what emotions are likely given the context (Carroll & Russell, 1996; Kayyal et al., 2015; Le Mau et al., 2021; Ong et al., 2015). The visual scene and the object someone is holding can produce dramatic categorical shifts in the interpreted meaning of expressions (Aviezer et al., 2008; Reschke et al., 2019; Righart & de Gelder, 2008a, 2008b). Explicit cues as to a person's mental contents can affect what emotion they appear to express (Wieser et al., 2014). In addition to direct contextual cues about events and mental states, interpretations are also affected by more abstract contexts, including familiarity (Baudouin et al., 2000), group identity (Elfenbein & Ambady, 2003), and experimental design (Barrett et al., 2019; Doyle et al., 2021; Hoemann et al., 2019; Lecker & Aviezer, 2021; Russell, 1994). Even individual differences in observers' emotion concepts can affect the processing of expressions (Brooks & Freeman, 2018; Brooks et al., 2018, 2019).

Contextual cues that induce strong emotion predictions can produce dramatic categorical shifts in how expressions are interpreted. A facial expression that is reliably judged to convey disgust can yield equally reliable judgments of anger when presented

with objects and gestures associated with anger (Aviezer et al., 2008, 2012a). Commonly, though, the influence of emotion predictions is not in suggesting a specific answer but in constraining the space of hypotheses. Partial or ambiguous information about the context ("it's the final point of a tennis game") can shift the predicted emotions while still allowing for very distinct possible situations ("she won the point" or "she lost the point") (Anzellotti et al., 2021). In Study 3, simply knowing the videos were drawn from the GoldenBalls gameshow shifted observers' interpretations, illustrating the role of probabilistic hypotheses in inferring emotions from expressions.

So, when seeking to understand a person's emotions in real-world social interactions, we argue that observers generate a hypothesis space of likely emotions given the context and then use perceived expressions to infer the posterior probability of emotional experiences. In Study 5, we implement a formal model of how this process supports social cognitive reasoning. In the Bayesian belief updating framework, people infer antecedent events by reasoning about which hypothetical situation was the most probable cause of the emotions a player appeared to experience. We show that by combining one group of observers' predictions of the emotions likely in each possible outcome with the emotion attributions a second group of observers made to each expression knowing the context of the game, we could capture the inferences that a third group of observers made about the situation that caused the expressions. We could predict the expressions and situations that would lead to accurate causal inferences and those that would lead to systematically below-chance performance. These results support the idea that conceptual knowledge plays a fundamental role in how observers make sense of expressions by shaping the space of hypotheses about what experiences are likely.

This formal model highlights a key difference between the views of *emotion recognition* and *emotion reasoning*. *Emotion recognition* frames emotion understanding as pattern matching where better access to richer perceptual expression information should enable more accurate social inferences. By contrast, we suggest that situating perceptual expression cues as information in an intuitive theory is a better model than extending the domain of *emotion recognition* to include dynamic and multi-

modal patterns of expressive behavior. In our view, understanding what subjective experiences are implied by observed expressions is an ill-posed problem that humans address by reasoning over an intuitive theory of other minds (de Melo et al., 2014; Ong et al., 2019; Saxe & Houlihan, 2017; Wu et al., 2018). Framing the interpretation of expressions as an ill-posed problem makes clear that some contexts[7] will induce structured priors under which certain expressions will be conditionally useful, whereas other contexts will induce priors that render the expressions uninformative or misleading. Therefore, including more and richer perceptual information can certainly make expressions more informative but does not make the expressions informative independent of context—expressions acquire relevance and meaning in relation to their role in a causal mental model.

There is enormous interest in building computer systems that emulate human emotion understanding. While the major focus has been modeling how observers interpret expressive behavior, our present work argues that expressions are a component of what will be required. A stimulus-computable model will need to capture how observers generate emotion predictions from event structure (Ong et al., 2015; Skerry & Saxe, 2015), infer emotions from expressions (Jack et al., 2014; Ong et al., 2021), and use this information in service of everyday social cognition, such as inferring what events someone is reacting to.

In the present work, we provide a formal model of how observers causally reason over latent emotion representations to infer events from expressions, but leave aside how observers arrive at emotion representations. To understand how observers make sense of expressions, future work should aim to model how expressions inform inferences of others' appraisals, as a function of inferences of their beliefs, desires, constraints, and plans, using a computational theory of mind (Baker et al., 2009, 2017). This then needs to be integrated with perceptual cues like facial and bodily expressions (Anzellotti et al., 2021).

Our results support the *emotion reasoning* account of human emotion understanding, which argues that observers use a causally structured intuitive theory of other

[7]Where context includes the experimental design, culture, stimuli, previous experience, etc.

minds. Perceptual information and conceptual knowledge mutually constrain which explanations are probable. We show that formally modeling these mutual constraints as inference in an intuitive theory can explain how observers infer which unseen events evoked others' expressions, and capture both what people get right about other minds, and what they get wrong.

3.8 Methods

Stimulus generation

We generated 88 silent videos of players who participated in the final round of the 5th season of the GoldenBalls gameshow, which aired in 2009[8]. Videos were 5-seconds in duration, depicting a single player's reaction in the moments leading up to, and following, the outcome of the Prisoner's Dilemma (PD). The first 2 seconds consisted of footage featuring the player immediately prior to both players revealing their decisions, and the remaining 3 seconds was composed of footage of the player reacting to the outcome. The raw footage was cropped and/or masked such that only a single player was visible in a given video. Overt cues as to the players' decisions were occluded. The broadcast footage regularly cuts between the two players, so the 5 seconds of footage was gathered by taking the scenes of the player most temporally proximal to the moment where the players simultaneously present their decisions to each other, the crowd, and the cameras.

The four outcomes were represented equally in the stimuli (N=22 for each outcome), reflecting the true distribution of play—across all 287 broadcast episodes, players cooperated 53% of the time (van den Assem et al., 2012), and the decisions of a player dyad (the two opposing players in a game) were statistically independent of each other (Burton-Chellew & West, 2012) (see Appendix A.2.1). The size of the jackpots were converted from British pounds to USD (1:1.533 USD:GBP) and scaled by 1.12 to adjust for inflation. In the subset of episodes used, jackpots ranged between

[8]Footage of the show was provided by Endemol UK.

the equivalent of $3.50 and $130,944 with a mean of $25,582.58 USD.

Empirical data collection

We collected all empirical data on Amazon Mechanical Turk (mTurk). Across all experiments, workers were not permitted to participate more than once. Thus, the number of responses contributed by workers is uniform within any dataset, and each dataset was generated by mutually exclusive workers. Financial compensation to workers targeted 12 USD/hr and was adjusted if the duration of an experiment was misestimated. Workers who began but were unable to finish an experiment (for instance, were unable to view the training video) were paid the full amount when they could be identified. We restricted workers by geolocation to the United States using mTurk credentials and asked workers to only begin the experiment if they were fluent in English.

In Studies 1, 3, and 4, (but not 2) participants were informed about the gameshow and the rules of the game. This included watching an example of a player dyad negotiating before revealing their decisions. The total sum of the jackpot was also presented, with British pounds converted to USD and adjusted for inflation to reflect contemporary value. In the example negotiation video, the announcer explains the rules of the game and the two players attempt to convince each other to cooperate (*Split*). The example video ends before the players reveal their decisions. These two players were not featured in 88 dynamic expression videos.

Every experiment included comprehension questions that were used to exclude workers. Participants were asked 3 forced-choice validation questions: an attention check following the example negotiation video and two comprehension checks for the meaning of emotion labels. In addition, participants supplied a free response as to whether they recognized any of the videos or images in the experiments. Participants were excluded if they failed the validation, recognized that the videos were from the GoldenBalls gameshow, or reported that they had technical difficulties that interfered with them completing the task properly. In Study 2, the example training video was not shown so there was no attention check question.

In Studies 1, 2, and 3, participants judged what emotions the players experienced based on different sources of information. All information provided to the participants was veridical. Participants judged the intensities of 20 emotions on continuous scales from *not at all* to *extremely*. Order of the emotion labels was randomized between participants. The 20 emotion labels were: *Annoyed, Apprehensive, Contemptuous, Content, Devastated, Disappointed, Disgusted, Embarrassed, Excited, Furious, Grateful, Guilty, Hopeful, Impressed, Jealous, Joyful, Proud, Relieved, Surprised, Terrified*.

In **Study 1**, we recruited N=195 participants on mTurk. Participants were informed about the context of the gameshow and viewed the example negotiation video. In 12 trials, participants were presented with the value of the jackpot, both players' decisions, the resulting payouts to the players, and asked to judge what emotions one of the players (the focal player) would experience. Participants were shown a static headshot of the focal player, which was taken from a frame in the 2 seconds of anticipation footage in the 5-second dynamic expression videos. Participants were informed that the photos were taken prior to the players revealing their decisions and were not photos of the players' reactions to the game outcomes. Of the participants who completed the experiment, 5 reported technical issues and 26 failed the comprehension questions. A total of N=164 (74 female, 4 unreported) were included in the analysis.

In **Study 2**, we recruited N=168 participants on mTurk. Participants viewed a subset of the dynamic expression videos and judged what emotions they thought the featured person was experiencing. In contrast to all other experiments, no contextual information about the videos was provided: participants were not told that the people shown were contestants on a gameshow nor told anything about the financial stakes. Each participant viewed 12 videos and provided continuous judgments of 20 emotions for each video. Of the participants who completed the experiment, 2 reported prior familiarity with the GoldenBalls gameshow and 30 failed the comprehension questions. A total of N=136 (55 female, 1 unreported) were included in the analysis.

In **Study 3**, we recruited N=168 participants on mTurk. Participants were informed about the context of the gameshow and viewed the example negotiation video.

In 12 trials, participants were presented with the value of the jackpot and a video of a player's dynamic expression, and asked to judge what emotions the player experienced. A static headshot of the player was visible while participants entered their emotion judgments. Participants were informed that the photos were taken prior to the players revealing their decisions and were not photos of the players' reactions to the game outcomes. Of the participants who completed the experiment, 10 reported prior familiarity with the GoldenBalls gameshow and 23 failed the comprehension questions. A total of N=135 (58 female, 4 unreported) were included in the analysis.

In **Study 4**, we recruited N=121 participants on mTurk. Participants were informed about the context of the gameshow and viewed the example negotiation video. Each participant responded to all 88 5-second expression videos. Dynamic expression videos played twice before any response could be made, after which participants could freely rewatch the videos. The total sum of the jackpot was also presented with each video and participants received visual feedback as to what payouts the players would receive based on the different combinations of choices. Before beginning the experiment, participants were incentivized with a bonus payment of 4 USD if they correctly inferred the outcome of at least 85% of the videos. The order of the videos was randomized between participants. Of the participants who completed the experiment, 1 reported prior familiarity with the GoldenBalls gameshow and 27 failed the comprehension questions. A total of N=93 (46 female, 3 unreported) were included in the analysis.

In **Study 5**, we used empirical emotion judgments to construct Kernel Density Estimates (KDEs) of population-level emotion distributions. In the full model, the weighted standard deviation of observers' emotion judgments determined the shape of the Gaussian kernel. For each emotion, the variance of the corresponding dimension of the kernel was calculated using Scott's Rule (Scott, 1992):

$$\sigma = s\, n^{-1/(d+4)} \tag{3.6}$$

where σ is the variance for a given dimension, s is the weighted sample standard

deviation of judgments of the corresponding emotion, n is the number of judgments, and d is the number of emotions. The kernel has zero covariance between the emotion dimensions, i.e. the off-diagonal variance-covariance matrix of equation (3.3) are all zero. Note that while the kernel is Gaussian and has zero covariance between the emotion dimensions, the structure of the data used to fit the KDE can induce multimodality and covariances between emotions. The KDE distribution of emotions attributed given expressions, $p(\boldsymbol{e}|x)$, was transformed to match the KDE distribution of emotions predicted given event context, $p(\boldsymbol{e}|a)$.

We measured model performance using Lin's Concordance Correlation Coefficient (*ccc*), which is a metric of the agreement between predictions and a gold standard or ground truth measure (Lin, 1989). Whereas Pearson's correlation is insensitive to whether samples differ in intercept or scale, Lin's concordance correlation penalizes deviations from the identity line (perfect prediction) making it a more stringent metric of explanatory power. Lin's concordance correlation gives the expected squared perpendicular deviation from a 45 degree line through the origin ($\mathrm{E}[(X-Y)^2]$), which is computed for a sample as,

$$ccc = \frac{2\, s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{3.7}$$

where s_{xy} is the covariance of x and y, s_x^2 is the variance of x and $\bar{x}$ is the mean of x.

Lesion Models. To be as generous as possible to the lesion models, we used grid search to fit the KDE kernel. As in the full model, a single bandwidth factor (b) scaled the empirical standard deviation of each emotion. For the lesion models, we chose the b that maximized the concordance of the data, where $\sigma = s\, b$. We fit the bandwidth factor for each lesion model using all of the data, without cross-validation or left-out test data, so the performance of the lesion models are likely to be inflated. Scott's factor ($b = n^{-1/(d+4)}$), which was the bandwidth factor used in the full model, was always included in the grid search. The lesion models follow the same procedure to transform the distribution of emotions judged from expressions: the

KDE distribution of each emotion was estimated using the given bandwidth, then the expression distribution was transformed to statistically match the probability density of the prediction distribution.

Classification of outcomes from emotions. In Studies 1, 2, and 3, we used linear support vector machines (SVM) to classify outcomes from observers' emotion judgments. The SVM were implemented in scikit-learn (Pedregosa et al., 2011) with the LinearSVC class, using a one-vs-rest scheme, squared hinge loss function, and ℓ_1 penalty. The emotion judgments corresponding to one stimulus corresponding to each of the four outcomes were held out of the training data. The SVM was trained to classify the outcome category from emotion judgments. One stimulus from each outcome was held out for testing. The hold-out procedure was iterated to test the out-of-sample classification performance for all of the data, using random sampling of which combination of stimuli were included in the test set.

For each test set, the SVM hyperparameters were fit to the training data using Bayesian optimization and k-fold cross-validation. The training data were resampled with replacement such that the number of emotion vectors was balanced across stimuli. The training data were centered and scaled with emotion by removing the mean and scaling to unit variance. The mean and variance of the test data were adjusted according to the training data.

Chapter 4

Generative model of inferred appraisals

> "And yet I have constructed in my mind a model city from which all possible cities can be deduced," Kublai said. "It contains everything corresponding to the norm. Since the cities that exist diverge in varying degree from the norm, I need only foresee the exceptions to the norm and calculate the most probable combinations."
>
> "I have also thought of a model city from which I deduce all the others," Marco answered. "It is a city made only of exceptions, exclusions, incongruities, contradictions. If such a city is the most improbable, by reducing the number of abnormal elements, we increase the probability that the city really exists. So I have only to subtract exceptions from my model, and in whatever direction I proceed, I will arrive at one of the cities which, always as an exception, exist. But I cannot force my operation beyond a certain limit: I would achieve cities too probable to be real."
>
> — Italo Calvino, *Invisible Cities*

4.1 Introduction

Human social life depends on our ability to understand, and critically, anticipate, other people's emotions. Intense efforts in both basic science and industrial applications are currently directed towards building models of emotion recognition: identifying the emotion that a person is or was feeling, typically based on facial or vocal

expressions. Here we tackle a complementary aspect of emotion understanding: predicting how a person will emotionally react to an event. Emotion prediction is critical as an input to planning in social interaction. People choose actions in order to cause some emotional reactions in their partners (with variable success). Thus, any model of human emotion understanding must incorporate the key capacity to predict how others will feel in response to events in social interactions.

To illustrate the phenomenon we target, imagine watching an episode of the popular British game show, 'Golden Balls' (van den Assem et al., 2012). During the episode, two players, Arthur and Bella, play a public one-shot social game called 'Split or Steal'. On the table is a pot of 100,000 USD. Eventually, each player will secretly choose to Split (`C`ooperate) or Steal (`D`efect). If both players choose to Split, each takes home \$50k. If both choose to Steal, they both leave with nothing. But if one chooses to Split and the other chooses to Steal, the one who stole takes the entire \$100k and the other player leaves with nothing[1]. Before Arthur and Belle make their choices, the game show host gives them a chance to talk to each other (in front of the live studio audience and TV viewers at home); they both vehemently promise to choose Split. Then they each make their secret choice. The choices are revealed simultaneously—they both chose Split! What do you predict Arthur will feel in this moment? Without seeing Arthur's reaction, human observers nevertheless generate systematic, specific predictions: for example, Arthur will feel *joy*, *relief*, and *gratitude*. By contrast, if Arthur split but Bella stole, observers predict he will feel *disappointment*, *envy*, and *contempt*.

The question for the current research is: How do human observers generate these emotion predictions? Social games offer a simple but evocative space of emotionally charged social interactions, that can be fully described in a simple set of quantitative inputs but afford diverse and fine-grained predictions about the players' emotions.

We develop a Bayesian framework to formalize the conceptual knowledge, and the

[1]This payoff structure is similar to a one-shot prisoner's dilemma (PD): the payoffs are symmetrical for the players, the players make their decisions without knowledge of the other player's choice, and it is never in either player's financial interest to cooperate. However, because both being defected on (`CD`) and mutual defection (`DD`) confer the same payoff (\$0), this game has been referred to as a "weak" PD (Rapoport, 1988).

inference process, that observers use to predict others' emotions in social interactions. Our model aims to capture three features of human emotion understanding: first, human observers make systematic, nuanced predictions of the emotional reactions others have to specific events; second, human observers tailor their emotion predictions to the specific player, knowing that individuals can react differently to the same event; and third, human observers make fine-grained predictions of when, and how much, people will experience distinctively social emotions—not just *joy* and *disappointment*, but also *gratitude*, *relief*, *envy*, and *guilt*.

To build a model of human emotion prediction that captures people's understanding of others' emotions, we integrate three key ideas about human social cognition into a single computational model. First, human observers predict a person's emotional reaction to an external event by inferring how that person is likely to interpret (or 'appraise') the event based on the person's mental state (including values, expectations, costs, etc., see Chapter 2; Figure 4-1). Thus, our model realizes a computational implementation of Appraisal Theory (Barrett, 2014; Ellsworth & Scherer, 2003; Moors et al., 2013; Ortony et al., 1990) for observers' third-party inferences. Second, human observers infer an individual's specific mental states from the person's actions, by assuming that his or her actions reflect approximately rational plans to maximize a subjective utility function. Our model thus extends models that take a Bayesian approach to Theory of Mind ($BToM$) (Baker et al., 2009, 2017). This extension enables our model to predict emotional reactions by way of inferences about the underlying preferences and beliefs which generated them. Third, observers know that people value outcomes that go beyond their own monetary gain; people also have higher-order social values, like equity and achieving a desirable reputation (De Bruyn & Bolton, 2008; Dufwenberg & Gneezy, 2000; Kleiman-Weiner et al., 2017). We hypothesize that observers use these non-monetary social values to predict social emotions. To capture emotion predictions that depend distinctively on social values, we incorporate weighted utility terms like equity and reputation into observers' mental model of how other people plan their actions. By inverting the planning model with these additional utility terms, the model jointly infers a player's social

and monetary preferences given the player's action in the game. It then combines these inferred social values with an external event (the opposing player's decision), to derive inferred appraisals. Inferred appraisal loading are then combined to yield quantitative predictions of emotion intensity judgments. This rich structure enables our model, like human observers, to predict a player's fine-grained emotional reactions to hypothetical events and tailor those predictions to individual players. Since the model aims to computationally recapitulate human emotion reasoning, we test if the model's emotion predictions match observers' predictions across a broad range of twenty nuanced emotions and twenty individual players.

The mental model that observers use to predict how others will experience events reflects observers' intuitive theory of other minds: a causally structured ontology of concepts comprised of people's lay knowledge (Gerstenberg & Tenenbaum, 2017; Lake et al., 2017), which is typically not explicit or fully introspectable (Murphy & Medin, 1985). Note that aim of this work is to build a formal scientific model of people's intuitive theory of emotion, not to test whether the intuitive theory is accurate. That is, although people are often able to both sensitively perceive and accurately predict others' emotions, in some contexts people also make systematic errors in emotion prediction, for example, over-estimating the duration or intensity of emotional reactions (Gilbert et al., 1998; Pollmann & Finkenauer, 2009). Because we are interested in capturing and characterizing people's intuitive causal theory, we do not here attempt to test the ground truth accuracy of either the observers' or the model's predictions, only their similarity to each other.

4.2 Background

To predict what emotions Arthur will experience when he splits the pot with Belle, observers employ an intuitive theory about the causal relationship between the events described and emotion concepts like *joy*. Analogous to appraisal theories of actual (first-person) emotions (Barrett, 2014; Ellsworth & Scherer, 2003; Moors et al., 2013; Ortony et al., 1990), we suggest that observers predict another person's emotional

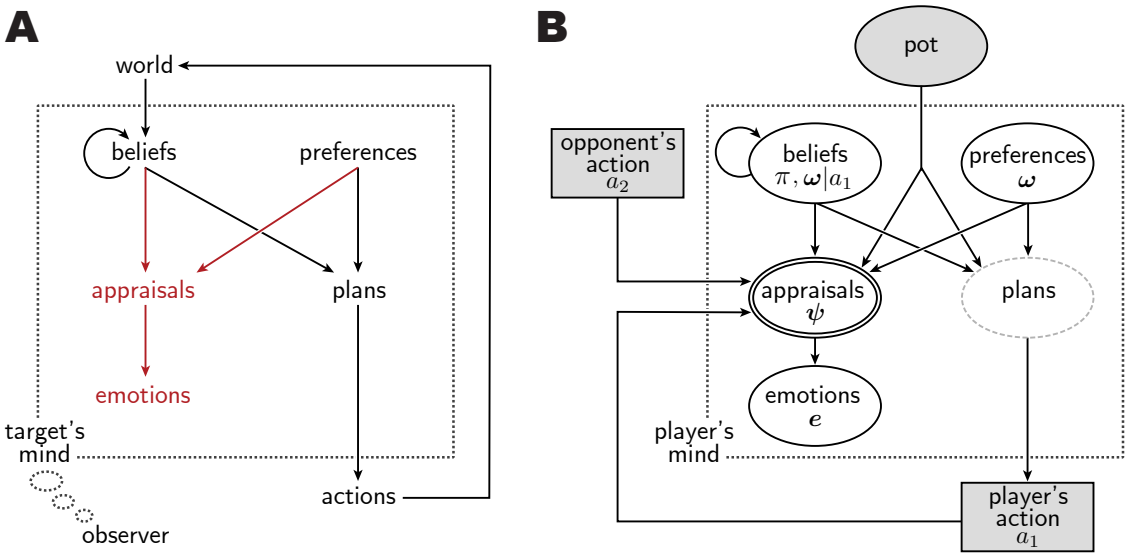

Figure 4-1: **Emotion prediction as inference in an intuitive theory of mind.** Hypotheses about how human observers reason about others' emotions can be formalized as probabilistic generative models. **(A)** The general inverse inference hypothesis about the causal relationship of a target's actions and emotions. We treat observers' emotion predictions as a function of their inference of how the players will subjectively evaluate, or "appraise" the game's outcome. Inferring a player's likely appraisals depends on an ill-posed inverse inference of how beliefs and preferences explain players' actions. This reflects a hypothesis about observers' lay theory of other people's minds, not a scientific hypothesis about people's actual emotions. **(B)** The implementation of the general hypothesis for a non-anonymous Prisoner's Dilemma. Here, observers predict a player's emotions by inferring what preferences and beliefs motivated the player to cooperate or defect, and reason about how those preferences and beliefs would cause the player to emotionally react to the outcome of the game. The intuitive theories we test take the form of directed acyclic graphs, where arrows indicate the causal relationship between variables. Shaded nodes are observable variables and open nodes are latent variables. Round nodes are continuous variables, rectangular nodes are discrete variables. Nodes with a single border are random variables. The double border indicates that appraisals are calculated deterministically. Plans are shown with a partial border because they are not explicitly represented in this model.

reactions by inferring how an external event relates to the person's internal mental states. This inference generates loadings on "appraisal variables", which include whether the situation is congruent with the person's goals and values, and consistent with their expectations (Scherer, 2005). Using these variables, simple classifiers can pick human observers' labels for many emotional events. For example, in one study using 6000 real-life events, a classifier given human ratings of an event in terms of 25 appraisal variables picked the correct emotion label (from 14 choices) for 51% of 6000 events (Scherer & Meuleman, 2013), see also (Skerry & Saxe, 2015). Thus, Appraisal

Theory offers a promising framework for capturing observers' third person emotion predictions (de Melo et al., 2014; Ong et al., 2015; Saxe & Houlihan, 2017; Skerry & Saxe, 2015; Van Kleef, 2010; Wondra & Ellsworth, 2015; Wu et al., 2018).

In contrast to earlier classifier models (Scherer & Meuleman, 2013; Skerry & Saxe, 2015), we aimed to make rational, situation-computable predictions for a player's emotional reaction to events. That is, rather than rely on human observers to manually annotate appraisal variables, we aimed to infer them directly from a description of the event itself, just as human observers do. Similarly, rather than rely on a researcher generated list of qualitative event features specific to a particular domain, we aimed to derive a more abstract space of appraisal variables that is more likely to apply across domains. In particular, the space of appraisal variables is grounded in utility: the accounting of the costs and benefits to oneself and others.

We were inspired by the success of a previous model: (Ong et al., 2015) studied how human observers predict the emotional reactions of others while participating in a lottery. Observers watched a single player spin a lottery wheel divided into three sectors of variable area indicating fixed monetary rewards of $25, $60 and $100, and predicted the player's emotional reaction to the outcome. Because the probability that the wheel lands on a reward is proportional to the area for that reward, each wheel configuration confers an expected value. The difference between how much a player won and the expected value of the wheel defines the reward prediction error. A model based on the amount the player won, the prediction error, and the absolute value of the prediction error, explained most of the reliable variance in observers' predictions of 8 emotions (happy, content, surprised, afraid, disgusted, angry, sad, and disappointed). For instance, the model (consistent with human observers) predicts that a player will feel more happiness, more surprise, and less disappointment, if he won $60 when he expected to win $30, than if he won $60 but expected to win $80.

While groundbreaking, (Ong et al., 2015)'s model has two major limitations. First, the event context is heavily constrained. A lottery can evoke *joy* and *disappointment*, but not emotions like *guilt* or *gratitude*, which depend on a player's intentions, choices, and social dynamics, since players make no decisions and have no social interactions.

Indeed, this model implicitly assumes that players' only goal is the maximization of their own monetary payoff. Yet it is well known that once a situation involves decisions that can impact other people, humans choose their actions to optimize additional values (Falk et al., 2003; Hayashi et al., 1999; Kiyonari et al., 2000; Melo et al., 2016). In a one-shot social dilemma like the 'Split or Steal' game described above, a selfish monetary utility maximizing agent would always choose 'Steal' (Rapoport, 1988). By contrast, real humans playing 'Split or Steal' chose to cooperate about half of time (van den Assem et al., 2012), suggesting that human players bring non-monetary social values into these games. Indeed, it is the tension between social values and the temptation to selfishly maximize monetary payoffs that makes social strategy games so compelling to play and to watch.

Unlike passive lotteries, predicting others' emotions in 'Split or Steal' is unlikely to be fully explained by monetary reward and reward prediction error alone. If Bella chooses 'Steal', for example, then Arthur's monetary payoff is independent of his choice (he leaves with nothing, regardless). Nevertheless, observers infer that Arthur likely has different values and expectations, and thus predict that he will experience different emotions, if Arthur chose 'Split' (e.g., more *envy*) versus if he chose 'Steal' (e.g., more *guilt*).

Furthermore, two observers might predict different emotions for the same person, if one observer has additional background knowledge of a specific individual to inform their inferences. That is, since appraisals are constructed from agent-relative expectations (in the form of priors and posteriors) background knowledge about an actor can influence the emotions predicted. Thus, based on prior knowledge, Bella's friends might infer that she chose 'Steal' only because she expected Arthur to steal, and not because she particularly values the money.

Our approach addresses these limitations. Our computational model of emotion prediction is based on a model of how observers infer players' specific values and expectations in social strategy games; we hypothesized that these inferences would provide the foundation for emotion predictions. We explicitly modeled how observers update their estimates of players' values and expectations from observations of the

players' actions in the game using an extension to $BToM$, and used those estimates to make emotion predictions. Then we tested whether this model could generalize to emotion predictions for a specific individual, when observers were given additional background knowledge that informed inferences about the individual's values and expectations.

4.3 Inverse planning with social values

How do human observers infer a specific person's preferences and beliefs? One clear source of information is the person's actions. People typically choose intentional actions that are likely to achieve their goals or maximize their rewards, given their expectations and beliefs (the principle of rational action). As a result, even a single sparse observation (e.g. observing one action) can lead observers to update estimates of the person's internal mental states such as beliefs, desires, costs, habits and intelligence, which cannot be directly observed (Baker et al., 2017; Evans et al., 2016; Gershman et al., 2016; Jara-Ettinger et al., 2016; Jern & Kemp, 2015; Jern et al., 2017; Kliemann & Adolphs, 2018; Kryven et al., 2016).

These inferences are well captured by 'Bayesian Theory of Mind' ($BToM$) models that probabilistically invert a forward model of rational planning. A forward model simulates how approximately rational agents, imbued with rich cognitive structure, perceive, plan, and act in a dynamic world. Probabilistic inversion of a forward model can then enable ill-posed inverse inferences of what preferences and beliefs were likely to have caused the observed behavior.

In our forward model we incorporate social equity utilities that account for people's actual decisions in social dilemmas (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999). Fehr and colleagues proposed that humans are motivated, to varying degrees, by two kinds of concerns for fairness in social interactions. Disadvantageous inequity aversion (DIA), a preference not to end up worse off than others, is a powerful and culturally conserved social preference (Blake et al., 2015; Henrich et al., 2005). In the context of 'Split or Steal', DIA is a preference not to be left with nothing while the other player

Belle (player 2)

Arthur (player 1)	C	D
C	\$50k, \$50k	0, \$100k
D	\$100k, 0	0, 0

Split or Steal

Money

a_1 \ a_2	C	D
C	1/2	0
D	1	0

AI

a_1 \ a_2	C	D
C	0	0
D	1	0

DI

a_1 \ a_2	C	D
C	0	1
D	0	0

Figure 4-2: **Payoff functions.** Decisions made by the two players jointly determine the players' relative payoffs in 'Split or Steal'. Payoff functions reflect how relevant the outcome is to a set of base values. The outcome's relevance is scaled by the size of the jackpot, which projects to outcome onto the dimensions of value, which are then weighted by player 1's preferences to yield subjective utilities; equation (4.1). With respect to monetary value, the payoff function simply returns the proportion of the jackpot that player 1 wins. When player 1 defects and the opponent cooperates, player 1 takes the whole pot: $Money(a_1{=}\texttt{D}, a_2{=}\texttt{C}) = 1{\cdot}pot$. Advantageous Inequity (AI) returns how much more player 1 received than player 2, and Disadvantageous Inequity (DI) returns how much more player 2 received than player 1. For the same decisions, $AI(\texttt{DC}) = 1{\cdot}pot$ and $DI(\texttt{DC}) = 0$.

Steals the whole pot. In addition, Fehr and colleagues observe that people's choices reflect advantageous inequity aversion (AIA), a preference not to extract more than one's fair share of a resource (Fehr & Schmidt, 1999). In the context of 'Split or Steal', AIA is a preference to share the pot with an opponent, rather than taking the whole pot oneself.

In using Fehr's parameterization of real choices as the basis for the intuitive theory of others' choices, we assume that observers have an intuitive grasp of the social motives that account for people's real-world behavior. Rational planning in the game thus maximizes utility over both non-social (monetary) and social (interpersonal inequity aversion) preferences, given expectations for the opponent's choice. Inverting this planning model would allow an observer to jointly infer a player's preference for selfish profit and inequity aversion, as well as the player's expectations for the

opponent's action, before the opponent's choice is known. Many combinations of values and expectations are consistent with each choice a player can make; nevertheless we propose that observers systematically change their estimates of player's values and expectations based on just a single observed action. So for example, given that Arthur chose 'Split', it is more likely that Arthur valued a good outcome for Belle (AIA) more than he valued avoiding exploitation (DIA), and more likely he expected Belle to choose 'Split'.

Once the opponent's action and the outcome of the game is known, the player's inferred values and expectations can then be used to infer the player's appraisal variables, including values (did she get what she wanted?), prediction errors (did she get what she expected?), and counterfactuals (how did this outcome compare to what could have happened?). For example, when the choices are revealed, observers predict Arthur will feel *joy* not only because he won more money than expected given his action (traditional reward prediction error), but also because Belle won money (better than expected with regards to AIA, which Arthur strongly prefers), and because this outcome is better than other outcomes that were possible. Including social values in inverse planning should thus allow the generative model to better fit people's predictions of social emotions, like *guilt* and *envy*, which are likely to depend mostly on a player's inferred social values and social prediction errors.

4.3.1 Modeling inverse planning with social values

We first simulate how players make decisions in an anonymous version of Split or Steal. In this Anonymous Game model, players have preferences exclusively for 'base' features, i.e. variables that are situation-computable. We use the three base features from Fehr and colleagues' parameterization of utility (Fehr & Schmidt, 1999): player 1's total monetary reward ($Money$), how much more player 1 received than player 2 (advantageous inequity, AI), and how much more player 2 received than player 1 (disadvantageous inequity, DI); see Figure 4-2. Not all players pursue the same values in these games; some players might be more motivated to maximize selfish gain, while others are more motivated to avoid unequal distri-

butions. To generate this variability across simulated players, we included non-negative preference weights (ω) that modulate the subjective utility players derive for rewards. In dyadic interactions like the Split or Steal game, player 1's utility is $U_1^{base} = \omega_{Money}^{base} Money - \omega_{AIA}^{base} AI - \omega_{DIA}^{base} DI$. The negative signs indicate that players seek to minimize inequity, thus ω_{AIA}^{base} and ω_{DIA}^{base} reflect a player's advantageous and disadvantageous inequity aversion. To adapt this utility function to simulate player 1's utility optimization under uncertainty, we express the utility features in terms of the players' possible actions, incorporate player 1's beliefs, and transform the utility features to accommodate a large range of possible rewards.

In the 'Split or Steal' game, players privately decide whether to split the pot by cooperating (`C`), or try to steal the pot by defecting (`D`), and reveal their decisions simultaneously. Payoffs are determined by the players' actions (a_1 and a_2 for player 1 and 2, respectively) and the pot size (Figure 4-2). The base features can therefore each be expressed a function of the tuple (a_1, a_2, pot). Simulated players are endowed with a weighted expectation about what choice their opponents will make (π_{a_2}), which models player 1's subjective belief about $\mathrm{P}(a_2)$. Players are also simulated as having a prior belief about the expected value of the game, π_{Money}. Human observers who participated in our studies were informed that pot sizes in the Split or Steal game can range from \$1 to over \$100K USD, and likely infer that contestants on a popular high-rewards game show have expectations about how much they could win. We therefore adjust monetary utility to reflect the difference between how much a simulated player expected to win before learning the pot size. Rewards that fall short of the reference point are perceived as negative utilities.

Since observers do not know a player's actual motivations, we simulate how players with different preferences and beliefs would act, then invert this forward simulation to infer the individual player's preferences and beliefs, given just a single observation of her action (e.g. $a_1 = \texttt{C}$). To test whether humans solve a similar inverse problem when reasoning about others' minds, we then compare the inverse inferences generated by this model to those of human observers. In the Anonymous Game model, a player is defined by the tuple over preferences for the base features and beliefs

$\langle\, \boldsymbol{\omega}^{base}, \pi_{a_2}, \pi_{Money}\,\rangle$. The expected utility of a player choosing a_1 is the product of the expected probability and the subjective utility of that decision, integrated over the outcomes possible given a_1:

$$\begin{aligned} &\mathrm{E}[U^{base}(a_1)] = \\ &\quad \sum_{a_2} \pi_{a_2} \cdot \Big[\omega^{base}_{Money} \cdot \nu\Big(\, Money - \pi_{Money}\Big) - \omega^{base}_{AIA} \cdot \nu\Big(AI\Big) - \omega^{base}_{DIA} \cdot \nu\Big(DI\Big)\Big] \end{aligned} \tag{4.1}$$

where $Money$, AI, and DI are functions of the tuple $\langle\, a_1, a_2, pot\,\rangle$, and $\nu(\cdot)$ applies a transformation commonly used in behavioral economics to account for people's diminishing marginal utility (Balaz et al., 2013; Tversky & Kahneman, 1992). The value function ν and the reference point π_{Money} reflect insights from prospect theory and the study of how people value uncertain rewards (Kahneman & Tversky, 1979). For this model of people's intuitive theory about others, we opt for a simple adaptation of the theory and do not fit the value function or reference point to any data. The ν transformation amounts to a sign-adjusted logarithm that treats gains and losses of utility symmetrically, and π_{Money} is sampled from a normal distribution with the mean fixed at 1,000 USD.

Simulated decisions follow probabilistically as samples from the softmax distribution of the expected utility: $\mathrm{P}(a_1 \,|\, \boldsymbol{\omega}^{base}, \pi_{a_2}) \propto \exp(\lambda \cdot \mathrm{E}[U^{base}(a_1)])$. The softmax decision function is a standard decision policy for modeling an agent's planning and decision-making in uncertain environments (see Luce, 1959) and for observers' reasoning about others' noisy choices (e.g. Baker et al., 2009, 2017; Evans et al., 2016; Jern et al., 2017; Kleiman-Weiner et al., 2017). The thermodynamic parameter λ determines how rationally vs. noisily decisions reflect differences in the expected utilities of the choices.

4.3.2 Comparison to human inverse planning

The first goal of our model is to capture the inferences observers make about players' preferences and beliefs. We therefore tested (a) whether human observers systematically infer the players' values and expectations from observing a single choice, and

(b) whether we could capture these inferences by inverting our generative model of players.

We presented Amazon mTurk participants with scenarios depicting one player's decision to cooperate or defect in an anonymous game, and asked them to judge how much the observed player cared about: *"acquiring as much money as possible"* (*Money*); *"sharing with the other player, not getting more than them"* (*AIA*); and *"having at least as much as the other player, not getting less than them"* (*DIA*). Observers also judged the player's belief about the opposing player (π_{a_2}), *"what did this person expect the other player to choose (and how confident is this person in their prediction about the other player's decision)?"*.

We find that human observers readily and consistently inferred the psychological base features from under-specified input (Anonymous Game, Figure 4-3). Returning to the example players Arthur and Belle from above, observers inferred that Arthur's (player 1's) decision to cooperate means that he was likely to be less motivated by *Money*, more averse to gaining an unequal and superior outcome (*AIA*), and less adverse to receiving an inferior outcome (*DIA*), than if he had defected. Observers also inferred that Arthur's cooperation strongly implied that he believed Belle was going to cooperate in kind. By contrast, if Arthur defected, observers were less confident about what Arthur expected Belle to choose, likely reflecting that observers found his choice consistent with multiple plausible explanations–he might defect in order to steal the entire pot or to avoid the sucker's payoff.

We then asked whether our model could capture these intuitive inferences. The Anonymous Game model simulates agents that generate softmax decisions, $a_1 \in \{\texttt{C}, \texttt{D}\}$, based on the expected utilities of the choices, with $\mathrm{E}\left[U^{base}(a_1)\right]$ being a function of the agent's base preferences, $\boldsymbol{\omega}^{base}$, and the agent's belief about what decision the other player will make, π_{a_2}. Since this generative formulation of play is invertible, using Bayes' rule we can infer the conditional joint distribution of a player's preferences and beliefs, given the player's decision:

$$\mathrm{P}\Big(\boldsymbol{\omega}, \pi_{a_2} \,|\, a_1\Big) \propto \mathrm{P}\Big(a_1 \,|\, \boldsymbol{\omega}, \pi_{a_2}\Big) \cdot \mathrm{P}\Big(\boldsymbol{\omega}, \pi_{a_2}\Big) \tag{4.2}$$

We ran the model in two ways to qualitatively and quantitatively test whether the structure of the model captures human inverse inferences. We first placed uninformative priors on the player's preferences and beliefs, which does not require fitting the priors to the data. Even without calibrated priors, the pattern of inferences inferred by the model captured the key qualitative patterns in the human data. We then replaced the uninformative priors with independent empirical priors computed from the human data (see BasePrior in Methods 4.7: Prior fitting). Player 1's preference weights and expectation of player 2's choice, as inferred by the empirically informed model, are shown in Figure 4-3 (Anonymous Game). This model closely matches human inferences of player's values.

In sum, a rational planning model can generate a plausible range of player choices, and can be inverted (like *BToM*) to make inferences about players' values from the observation of a single choice they made. Observers appear to make similar systematic inferences of players' values from the same observation. Both observers and our models are thus successfully resolving an ill-posed inverse problem, to recover four different mental states (values for selfish and social outcomes, and expectation for the other player) from the observation of players' behavior.

4.3.3 Second-order preferences: players' motive to enhance their reputation

The Anonymous Game model is missing a critical element of social strategy games: players' motive to enhance their reputation. In addition to preferences for certain kinds of outcomes, players have preferences for how they will be perceived. For example, Arthur may choose to cooperate primarily to signal his cooperativeness to future social partners. We hypothesize that human observers also infer these second-order preferences. Most importantly, for current purposes, inferences about the motive to enhance one's reputation likely underlie the prediction of key social emotions, like *pride* and *embarrassment*. To capture observers' reasoning about players decisions in a highly public context, we extended the generative model of decision planning to

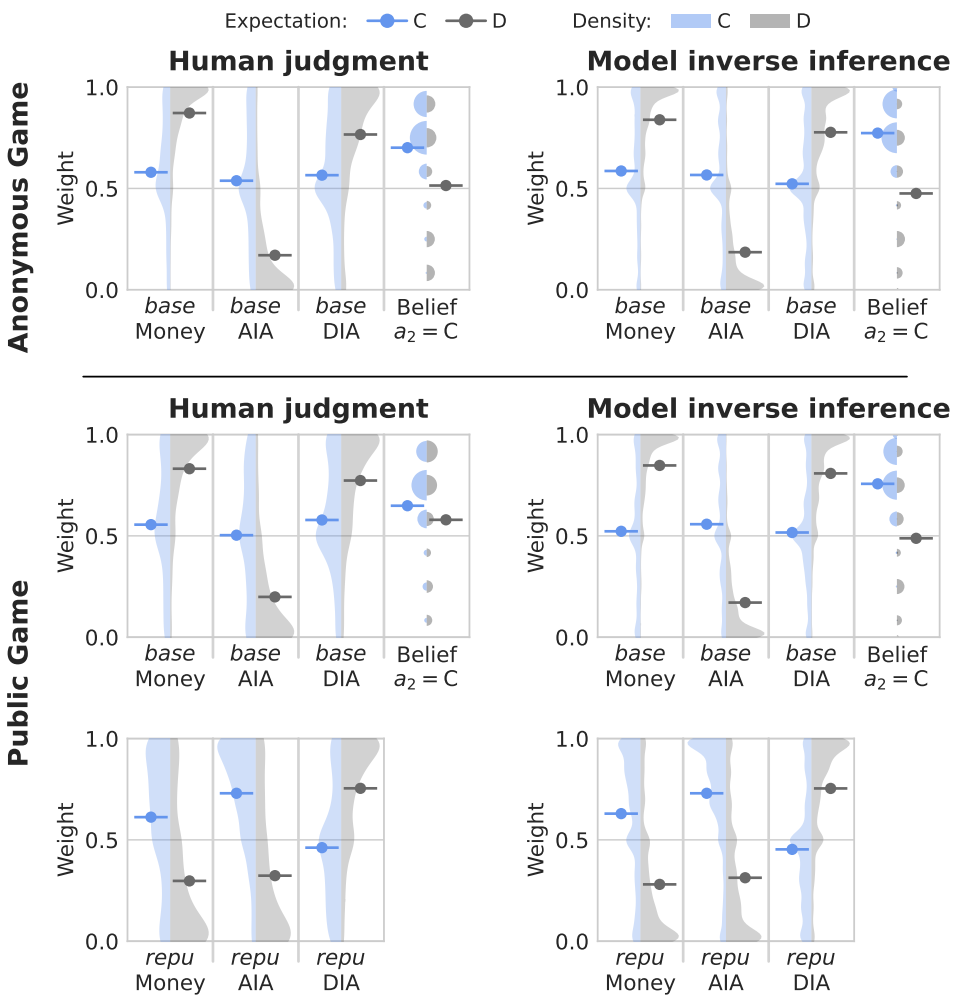

Figure 4-3: **Inverse planning.** Human observers were shown a player's decision to cooperate (C) or defect (D) and judged the player's likely preference and belief values. Preference weights take continuous values between zero and one. Player 1's belief about what player 2 will choose was rated on a 6-point confidence scale. Observers judged the player's three base preference weights and belief about player 2's intended action, $\pi(a_2{=}\texttt{C})$. In the Public Game, observers additionally judged the player's reputation preference weights. Point estimates give the expectation of each marginal distribution, conditioned on a_1.

include players' reputation concerns.

A standard way to incorporate reputation concerns might be to add additional base utility components that define what reputation signals players expect of their actions, which are not directly situation-computable and must therefore be specified for each situation and action. We follow a more cognitively natural strategy, whereby players apply their own theory of mind to anticipate how others will evaluate them

(Kleiman-Weiner et al., 2017). To choose an action that is reputation enhancing, a player must first infer how that action will be perceived by others. This requires an embedded inference loop over the base features (*Money*, *AIA*, and *DIA*), which are objectively defined by the event. We model reputation as a function of the inferences a rational observer would make about the weights of a player's base utility function. Each of these inferences are themselves weighted and treated as "second-order" utilities: an agent's preference for certain inferences that others would make about their values.

For each feature we introduce a reputation consideration term consisting of a weighted preference, ω^{repu}, and an expectation of what other people would infer about the agent given her action, a_1:

$$\begin{aligned} \mathrm{E}\left[U^{base+repu}(a_1)\right] = \\ \mathrm{E}\left[U^{base}(a_1)\right] - \omega^{repu}_{Money}\cdot\nu\Big(\mathrm{E}\left[\omega^{base}_{Money} \mid a_1\right]\cdot pot\Big) \\ + \omega^{repu}_{AIA}\cdot\nu\Big(\mathrm{E}\left[\omega^{base}_{AIA} \mid a_1\right]\cdot pot\Big) \\ + \omega^{repu}_{DIA}\cdot\nu\Big(\mathrm{E}\left[\omega^{base}_{DIA} \mid a_1\right]\cdot pot\Big) \end{aligned} \tag{4.3}$$

where $\mathrm{E}\left[\omega^{base} \mid a_1\right]$ are the expectations of the preference weights inferred by the Anonymous Game model, shown in Figure 4-3.

Note that the sign on the reputation utilities is opposite that of the base utilities, reflecting the prediction that human observers believe that players desire to be perceived as motivated to improve equality, and not motivated to selfishly maximize their own monetary payoffs. Expected reputation utility is scaled by the pot size and combined with the expected base utility to give an expected utility for each action. Agents simulated by the Public Game model make probabilistic decisions from the softmax of the expected utility, $\mathrm{E}\left[U^{b+r}(a_1)\right]$.

First, we simulated players with both base and reputation values. We find that these players are more likely to cooperate than players with base values alone. Simulated players with only base values defected on 59% of trials, whereas simulated players with both base values and reputation concerns defected in 48% of trials. Ac-

tual humans playing 'Split or Steal' for real stakes in a televised game show defected 47% of the time (van den Assem et al., 2012).

Next, we modeled observers who see only a player's choice, and infer both base and reputational values simultaneously. Inverting the Public Game model gives a joint inference of the distributions of 7 random variables conditioned on the agent's decision: 3 weights for the base preference of each feature, 3 weights for the reputation preference of each feature, and one expectation for the other player's action. We again presented online participants with scenarios depicting one player's decision, this time in the highly public context of the Spit or Steal game show. The observers judged players' base and reputation preferences, and belief about the other player's intended decision. Observers' judgments and the inverse inference of the Public Game model are shown in Figure 4-3.

As a generative model of decision making, the Public Game model is much richer than is necessary to predict players' choices in a Prisoner's Dilemma, which can be captured by extremely simple models (Sally, 1995). Nevertheless, our interest is in what observers infer about the latent mental contents that underlie the decisions made by others. For this purpose, the richer model better captures the inferences human observers make and supports the subsequent inference of players' fine-grained reactions. Most importantly for current purposes, we expect that this richness is necessary to capture the predictions that observers make about players' emotions.

4.4 Emotion predictions

Through the successive inversion of increasingly rich generative models of behavior, we have built a model that uses a player's choice in a social game to infer the joint posterior probability of the player's selfish, social, and reputational preferences and belief about the opposing player's intended action. We now return to the challenge that we began with: testing whether our computational model can capture the conceptual knowledge and intuitive reasoning that underlie human observers' emotion predictions.

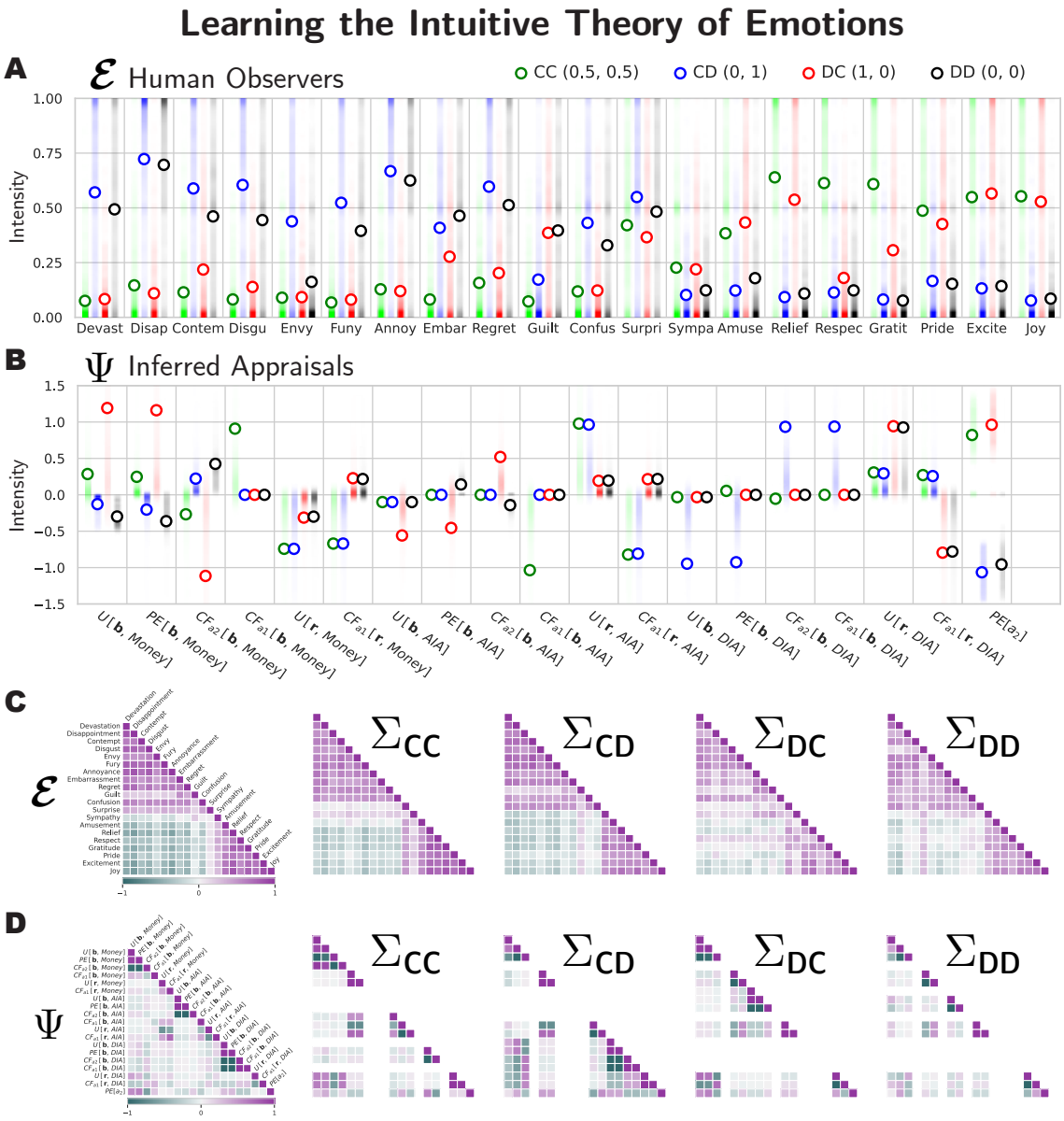

Figure 4-4: **Generative structure.** **(A)** Emotion predictions for the *GenericPlayers*. Circles show the expected intensity for each outcome, summing over pot sizes and the 8 photos. Shading shows the density of judgments. Color indicates the outcome of the games. $\boldsymbol{\mathcal{E}}$ is the matrix of the 20-dimensional emotion prediction vectors. **(B)** Expectations and densities of the normalized inferred appraisals. $\boldsymbol{\Psi}$ is the matrix of the 19-dimensional appraisal vectors. **(C, D)** The leftmost correlation matrices shows the Pearson's r for $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\Psi}$, respectively, which reflect between-outcome, between-pot, and within-stimulus covariance. The following matrices show the average within-stimulus correlation: a correlation matrix was calculated for every stimulus and then averaged within outcome.

4.4.1 Human observers' emotion predictions

We collected human observers' predictions of the emotions players would experience when the outcome was revealed in 'Split or Steal' games. We collected two data sets from online participants. The training data (n=554) was used to learn a transformation between the latent space of the inverse model and the emotion predictions, which was then used to predict emotions for the test data (n=1512). In the test data, observers were presented with specific information about each focal player. Collection of these data, referred to as the *SpecificPlayers*, will be described in Section 4.5.

In the training data (*GenericPlayers*), observers were briefed on the structure of the Split or Steal game and watched a video taken from the show in which the presenter explains the rules and two players negotiate in an attempt to convince the other to choose 'Split' (*cheap talk* negotiation). The introductory video ends before the players reveal their choices. Observers completed 8 trials, in which they saw a photograph of the focal player (designated player 1), a pot size (ranging from 2 USD to 207,365 USD), and the actions chosen by both players in that game. Observers saw 2 games for each category of payoff: `CC` where both players cooperated and each won half; `CD` where the player 1 cooperated and received nothing; `DC` where player 1 defected and took everything; and `DD` where both players defected and both got nothing. Observers then predicted how much player 1 would experience 20 different emotions: *Devastation*, *Disappointment*, *Contempt*, *Disgust*, *Envy*, *Fury*, *Annoyance*, *Embarrassment*, *Regret*, *Guilt*, *Confusion*, *Surprise*, *Sympathy*, *Amusement*, *Relief*, *Respect*, *Gratitude*, *Pride*, *Excitement*, and *Joy*.

To learn a transformation between the model and human emotion judgments, we make use of the rich structure present in the observers' emotion predictions. The *GenericPlayers* data (Figure 4-4A) illustrate that, even from sparse event depictions, human observers made systematically different emotion predictions for players in the four different types of payoffs. At the coarsest qualitative level, observers predicted that players who won money (`CC` and `DC` outcomes) would experience more positive emotions and players leaving with nothing (`CD` and `DD` outcomes) would experience

more negative emotions (Figure 4-4A). However, observers' emotion predictions do not just reflect the monetary outcomes of the game. For example, when player 2 defects, player 1 necessarily receives no monetary reward, yet observers predicted that player 1 would have different emotional reactions depending on whether he chose to cooperate (more *envy* and *contempt*) versus defect (more *guilt*). Note that preference/belief attributions and emotion attributions were collected from mutually exclusive groups to avoid cueing observers to think about emotion attributions in terms of the planning variables or vice versa.

4.4.2 Learning the latent structure of the intuitive theory of emotions

We hypothesized that human observers infer players' values and expectations from their actions using inverse planning, and predict players' emotional reactions to events based on those inferred mental states. In this model, emotion predictions reflect the observers' inferences about how players would react to an event given their particular beliefs and preferences. Critically, we assume that emotion prediction relies on inverse planning. Mental contents inferred via the inversion of an intuitive theory of mind generate the basis for inferring how a player will evaluate events. We use *appraisal* to refer to the inferred process of emotion generation, and thus call our model of emotion prediction based on mental states inferred from inverse planning, an Inferred Appraisals model. First, the model uses the pot size and player 1's chosen action to update estimates of player 1's preferences and beliefs. These preferences and beliefs then affect how the player reacts to (appraises) the situation. To generate inferred appraisals for the *GenericPlayers*, we ran the Public Game model in the same way as Section 4.3.3: using the empirically-derived prior over the base preferences and beliefs, we inverted the hierarchical generative model of behavior. Then, we estimated how an agent would appraise the outcome of the game. Appraisals are derived as subjective utilities, prediction errors, and counterfactuals, on a player's beliefs, base preferences, and reputation preferences. Methods 4.7 and Appendix B.1.4 detail how appraisals

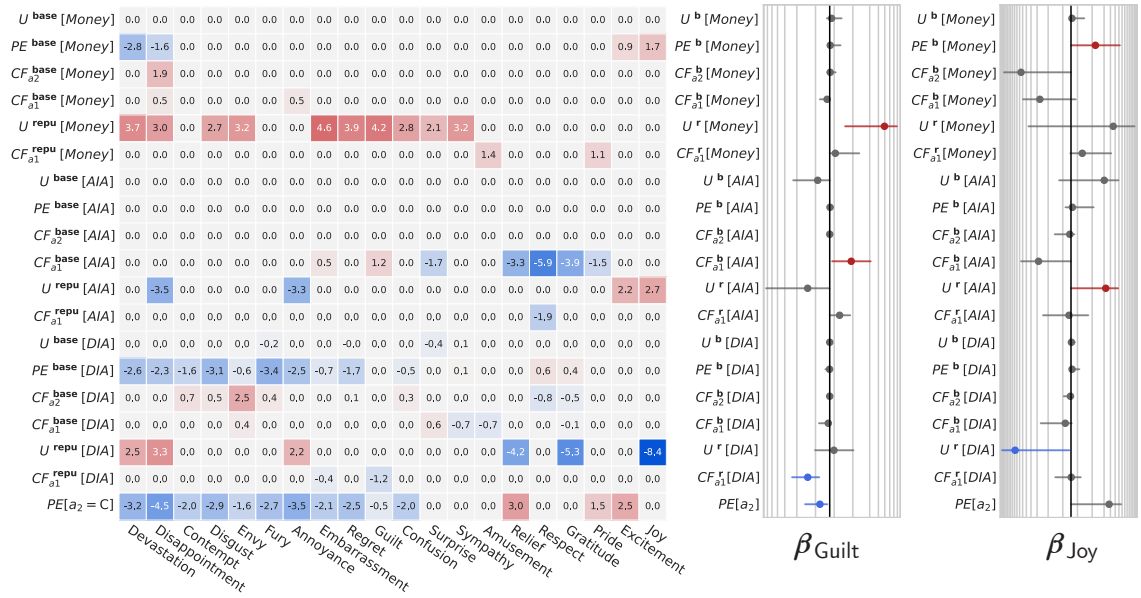

Figure 4-5: **Appraisal structure of the intuitive theory of emotions.** The β weights of the transformation were learned based on the joint distribution of appraisals and the joint distribution of emotion predictions for the *GenericPlayers*. To determine the scale of the Laplace prior, which induces a sparse solution, we cross-validated on subsets of the *SpecificPlayers*. (left) Mean weights of the learned transformation. (right) Example emotions glosses. We used gradient descent to discover modes, then variational inference to approximate the joint posterior distribution over β weights. These log-scale plots show the expectation and 95% HPD (Highest Probability Density) interval of the weights for *guilt* and *joy*. Color indicates that the HPD is above (red) or below (blue) zero.

are computed.

In order to generate specific emotion predictions, we need to learn the 'meaning' of each of the 20 emotion labels that human observers rated, in terms of the set of appraisal variables. There are many possible ways to accomplish this step (essentially, writing a dictionary of emotion labels in terms of the inferred appraisals computed by our model). It might be possible to constrain these definitions manually, by consulting formal and intuitive theories of the meanings of these emotion labels (Battigalli & Dufwenberg, 2007; Battigalli et al., 2015; Houlihan et al., 2018; Scherer & Meuleman, 2013; Sell et al., 2017; Sznycer, 2019; Sznycer et al., 2021; van Baar et al., 2019), but here we preferred to learn the transformation from appraisals to emotions.

To learn the function relating emotion labels to inferred appraisals, we made a strong assumption about the generative structure of observers' emotion predictions:

when people are asked to predict a player's emotions, they do not make twenty independent inferences but rather infer a joint distribution over the player's preferences and beliefs, and reason about how these inferred mental contents would cause the player to evaluate the situation (Figure 4-1). For instance, if player 1 cooperated while player 2 defected (`CD`), how much an observer thinks that player 1 wanted to avoid being disadvantaged will be reflected in that observer's predictions of player 1's experience of both *embarrassment* and *envy*. Thus, the covariance patterns of observers' emotion predictions reflect the latent structure of their intuitive theory of psychology. We make use of this information to learn a mapping between people's empirical emotion predictions, found in the training data, and the joint distribution over appraisal variables generated by the model.

Figure 4-4 shows the emotion predictions and inferred appraisals for the *GenericPlayers*. Figure 4-4A and B show the expected values by outcome. Figure 4-4C and D each show the overall covariance, and the average within-stimulus covariance for each outcome. Note that formalizing inferred appraisals as a probabilistic generative model permits us to leverage within-stimulus covariance in latent structure discovery.

Based on the *GenericPlayers* data shown in Figure 4-4, we learn a sparse transformation between the joint distribution of inferred appraisals and the joint distribution of emotion predictions. Specifically, we treat the empirical vectors of emotion predictions as observations from some function of the posterior distribution of inferred appraisals. We find a transformation of the appraisal distribution that maximizes the probability of observing the empirical data under a Laplace prior on the transformation coefficients. This yields a sparse linear transformation between inferred appraisals sampled from the Inferred Appraisals model and continuous quantitative predictions for the player's emotions (Figure 4-5).

Thus, the Inferred Appraisals model takes priors on players' preferences and beliefs, inverts a cognitively structured generative model of behavior to infer a joint distribution over inferred appraisals and transforms the distribution into emotion predictions. This model of human emotion understanding generates predictions of the emotions that human observers will attribute to players for arbitrary games (de-

scribed by the actions of the two players in a dyad and the size of the jackpot). The emotion glosses shown in Figure 4-5 reflect a computational hypothesis about the intuitive theory of emotion. For example, when observers predict that a player will experience *guilt*, this hypothesis says that the intensity prediction reflects an intuitive computation about the players' mental contents. Specifically, the player will experience more guilt when the player values his reputation of not being motivated by selfish monetary gain, but believes that others will think he is ($U^{repu}\, Money$). The player will experience more guilt if he cares about his opponent's welfare and could have made a decision that would not have put his opponent in a disadvantageous position, but chose not to ($CF_{a_1}^{base} AIA$). The player will experience less guilt about making a decision to take advantage of his opponent if he values being seen as a fierce competitor ($CF_{a_1}^{repu} DIA$). And the player will experience less guilt if his opponent tried to steal the jackpot ($PE\pi_{a_2=\mathtt{C}}$). We next test if the computational hypothesis formalized by the Inferred Appraisals captures human emotion predictions.

4.4.3 Comparing the Inferred Appraisals model to human observers

The Inferred Appraisals model generates a joint distribution over 20 emotions based on a pot size, the two players' actions (a_1, a_2), and the prior $\mathrm{P}(\boldsymbol{\omega}, \pi_{a_2})$. Using the transformation we learned based on the *GenericPlayers*, we generated emotion predictions for the *SpecificPlayers* (described in detail in Section 4.5). The Inferred Appraisals model captures the overall pattern of human emotion judgments. Positive emotions are predicted when players win money and negative emotions when players lose money. In addition, the model captures some of the more nuanced features of the empirical judgments. For example, consider the player shown in Figure 4-6. When player 2 defects, causing her to leave with no money, the model (like human observers) predicts more *envy* if she cooperated. When player 2 cooperates causing her to win money, the model (like human observers) predicts more *gratitude* and *respect* if she cooperated.

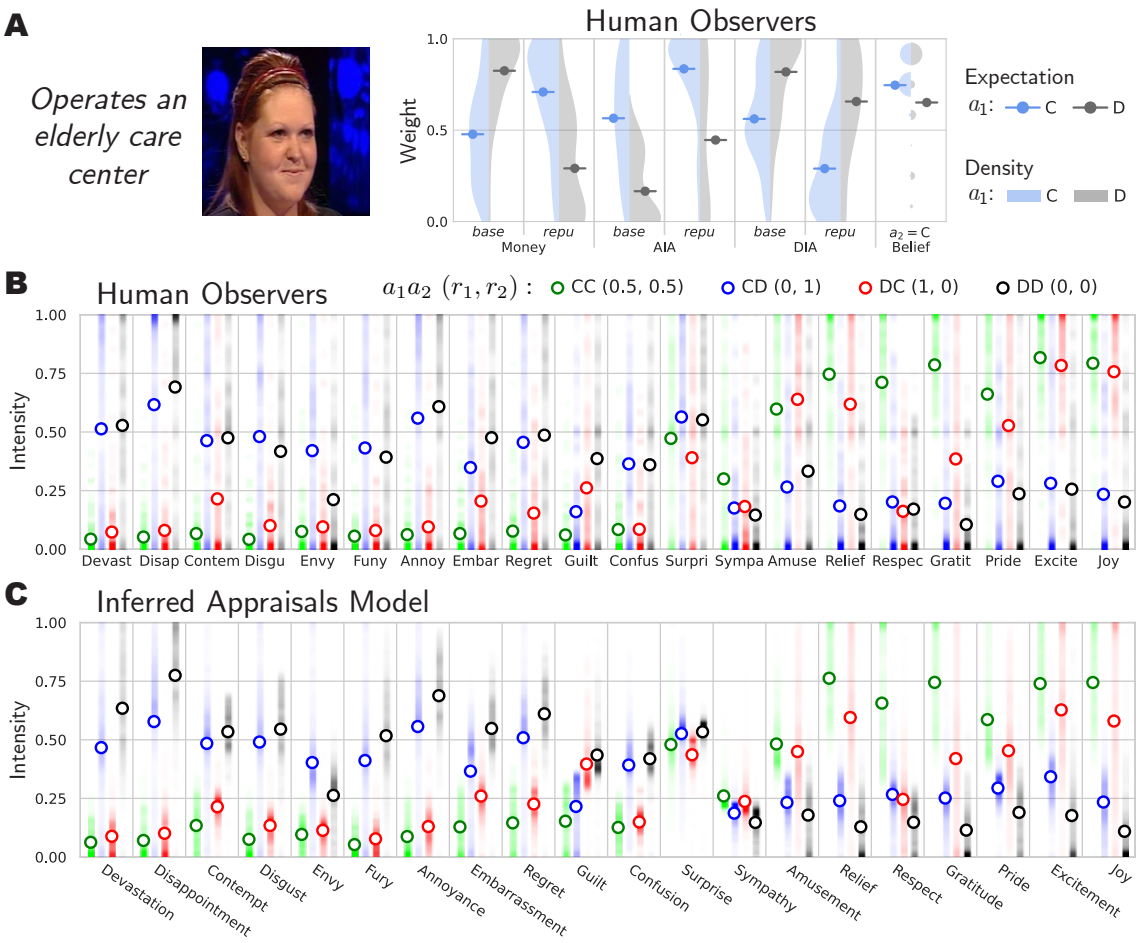

Figure 4-6: **Personalized inferences.** Observers were given personalizing information and a photo of a *SpecificPlayer*. These predicted that these abstract cues would induce priors over mental state inferences that would translate into biases in emotion judgments. **(A)** Human judgments of an example *SpecificPlayer*'s preferences and belief, marginalized over pot sizes. **(B)** Human judgments of the example player's emotions, made by an independent group of observers. Circles show the expected intensity for each outcome, summed over pot sizes. Shading shows the density of judgments. Color indicates the outcome of the games. **(C)** Personalized emotion predictions generated by the Inferred Appraisals model. The model simulated emotion predictions for the example player based on the preference and belief judgments shown in (A).

To assess the explanatory power of the model, we used Lin's Concordance Correlation Coefficient (*ccc*), which is a metric of the agreement between predictions and a ground truth measure (Lin, 1989). Whereas Pearson's correlation is insensitive to whether samples differ in intercept or scale, the *ccc* penalizes the model's deviations from human judgements by the mean squared error. Across all players, emotions, outcomes, and pot sizes, the Inferred Appraisals model fit the observer's emotion

predictions for the *SpecificPlayers* data well: $ccc = 0.869$ [0.855, 0.872].

Predictions of different emotions depend on different types of information, making it likely that a model's latent representations will enable it to capture some emotions better than others. Similarly, human observers can find a stimulus ambiguous with regard to one emotion but unambiguous with regard to another. Figure 4-8B shows the reliably of observers' predictions and how well the Inferred Appraisals model captures the empirical emotion predictions for the *SpecificPlayers*. To test whether the rich generative structure of the Inferred Appraisals model significantly contributed to its ability to capture observers' emotion predictions in this task, we compared the Inferred Appraisals model with two simpler alternatives.

4.4.4 Inverse planning lesion model

The Inverse Planning lesion selectively blocks inverse planning by inferring appraisal variables based on the prior distribution of beliefs and preferences (before any player acts), rather than the posterior distribution (based on the player's choices). Thus, the Inverse Planning lesion model generates emotion predictions using exactly the same rich social features with informative priors as the Inferred Appraisals, but does not update posterior estimates based on players' actions (see Figure 4-7).

This lesion model tests the importance of the shape of the joint distribution of appraisals. For example, if observers intuit the mental contents that would cause a player to feel *guilty* about defecting are also involved in planning what action to take, then observers might further infer that the players who would feel more *guilt* after defecting are less likely to decide to defect. The full Inferred Appraisals model uses the posterior distribution, $\mathrm{P}(\boldsymbol{\omega}, \pi \mid a_1)$, to derive appraisals which induces a conditional dependence between planning and emotion prediction, whereas the Inverse Planning lesion prevents updating the inference of a player's beliefs and preferences based on observing her action, a_1. Without a causal link to behavior, the inverse inference of players likely preferences and beliefs, reduces to the prior. Thus, the posterior probability in equation (4.2) becomes: $\mathrm{P}(\boldsymbol{\omega}, \pi_{a_2} \mid a_1) = \mathrm{P}(\boldsymbol{\omega}, \pi_{a_2})$. We similarly lesion the embedded inverse planning loop, which simulates how infer their behavior will

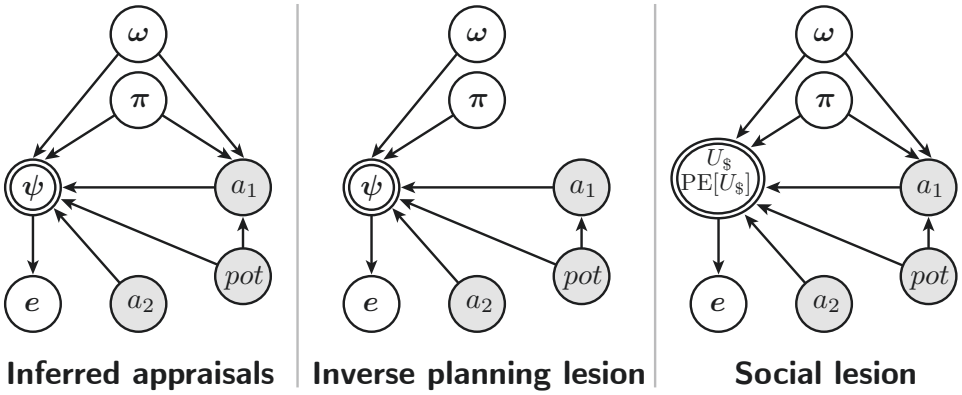

Figure 4-7: **Structure of lesion models.** The notational schema is identical to Figure 4-1; the full Inferred Appraisals model is reproduced in the first panel. The Inverse Planning lesion (middle) severs the causal link between simulated agents' mental contents and decisions in the game. Without the causal link, the inverse inference of players likely preferences and beliefs reduces to the prior. Appraisals ϕ still depend on a_1. The Social lesion (right) allows simulated agents to plan and act identically to the Inferred Appraisals model using social and reputational motivations. However, predicted emotions depend only on two inferred appraisals, monetary utility, and reward prediction error on monetary utility.

be interpreted by others. Appraisal generation is identical to the Inferred Appraisals model and still depends heavily on a_1.

To illustrate the Inverse Planning lesion, consider the effect of agents' beliefs about their opponents' actions (π_{a_2}) on the appraisals made by each model. In the full Inferred Appraisals model, as for human observers, simulated agents only tend to cooperate when they believe their opponent is also going to cooperate (see $\mathrm{E}[\pi_{a_2} \mid a_1{=}\texttt{C}]$ in Public Game of Figure 4-3). Inverse Planning lesion model agents cannot select actions based on their mental contents so the expectation of monetary utility reflects the prior on belief, $\mathrm{P}(\pi_{a_2})$, rather than what beliefs where likely given the agent's action, $\mathrm{P}(\pi_{a_2} \mid a_1)$. Thus, players simulated by the Inverse Planning lesion model end up in situations that they would not self-select into if they had intentional agency. These effects are relatively minor overall ($ccc = 0.836$ $[0.835, 0.836]$; Figure 4-8C), but are evident in specific emotions. For instance, the Inverse Planning lesion caused notable decrements in the capture of *envy*, *fury*, *guilt*, *surprise*, *respect*, and *gratitude* (Figure 4-8B).

4.4.5 Social lesion model

Whereas the Inferred Appraisals model infers appraisals that are functions of social equity and second-order (reputation) preferences, which we found to capture predictions of emotions like *envy*, *embarrassment*, *guilt*, and *joy*, in the Social lesion model we removed all of the social (non-monetary) values attributed to players. This lesion allows us to test the importance of these social values for successfully generating human-like emotion predictions.

The Social lesion model leaves forward-planning intact: generated game play, inferred monetary utility and prediction error are all identical to the full Inferred Appraisals model (see Figure 4-7). That is, in the Social lesion model, simulated agents plan and act according to the same policy as in the Inferred Appraisals model, but only two inferred appraisals are used as features for emotion prediction. This can be likened to observers having an intuitive theory that players' behavior depends on social considerations but their emotional reactions depend only on their monetary evaluations. The Social lesion model predicts 20 emotions from the transformation of the joint distribution of monetary utility and monetary prediction error, $\mathrm{P}(U^{base}_{Money}, PE^{base}_{Money})$. While low dimensional, this joint appraisal distribution is still a richly structured posterior from an inverse planning model with informative priors.

Across all players, combinations of decisions, and pot sizes, the Social lesion model showed a lower fit to human observers' emotion predictions: $ccc = 0.641$ [0.640, 0.641] (Figure 4-8C). The poor fit of the Social lesion model to the current data contrasts with prior research. In a lottery, how much money players won, and how much they should have expected to win, explained the majority of observers' predictions of eight emotions (Ong et al., 2015). In the context of the highly social Split or Steal game, social values were required to capture the patterns of human observers' predictions, including for the emotions that overlap between these studies (happy/joy, disgusted, angry/furious, sad, disappointed), Figure 4-8B.

For example, observers' predictions of players' *joy* in this game have a positive relationship with the pot size: players who win 13k USD by cooperating with their

opponents are expected to experience more *joy* than if they win 6k USD by cooperating. However, for any given pot size, observers predict approximately equivalent intensities of *joy* for players who defected and took the entire jackpot, as for players who cooperated and took half as much. This is true for large pot sizes, where half the pot could be over 100,000 USD, as well as for small pot sizes, where the difference could be tens or hundreds of dollars. Across games with jackpots spanning five orders of magnitude, the expectation of observers' predictions of *joy* have a positive relationship with the pot size but the expectation of *joy* given mutual cooperation (where player 1 wins half the pot) does not reliably diverge from the expectation of *joy* given successful defection (where player 1 wins the whole pot). With access to only monetary appraisals, the Social lesion can capture the positive relationship between *joy* and the pot size, but not the way in which *joy* seems to depend on a player's choice. With access to a broader set of appraisals, the Inferred Appraisals model infers that *joy* additionally depends on social evaluations, which includes wanting to have a reputation for not taking advantage of one's opponent. Modeling the generative structure of the intuitive theory enables us to discover these latent computations directly from observers' emotion predictions. Figure 4-5 shows the inferred appraisal structure for *joy* reflects observers' latent inference that the players won more money than they expected to (PE^{base}_{Money}) and enhanced their reputations for being considerate (U^{repu}_{AIA}).

The emotion best fit by the Social lesion model in the Split or Steal context was *disappointment*. Human observers predicted that players would experience *disappointment* when they do not win money, and the expected intensity does not vary considerably by pot sizes, with expectations ranging from 0.6 to 0.9 for the `CD` and `DD` payoffs. Observers also predict that players will experience *disappointment* when they win small sums (i.e. the value of the game was low to start with). The Social lesion model fit these prediction patterns because the model's latent representations are based on inverse planning of economic choices, informed by prospect theory.

By contrast, the Social lesion was largely unable to capture predictions of social emotions like *envy*, *guilt*, and *respect*. Observers only systematically expect *envy* in the `CD` payoff, where player 1 receives nothing and her opponent takes the entire

jackpot. This is a difficult pattern to represent in terms of monetary reward since player 1 also receives nothing in the mutual defection payoff (DD) but is not expected to experience *envy*. Similarly, *guilt* is reliably predicted for players who defect and is largely insensitive to whether the outcome of player 1's decision to defect is that she takes home the entire jackpot (DC), or nothing at all (DD). In sum, comparing the full Inferred Appraisals model to the Social lesion model reveals the importance of social values like inequity aversion and reputation concerns, in the structure of human observers' inferred appraisals and emotion predictions. Next, we tested a second assumption of our Inferred Appraisals model: that human observers use the player's own inferred beliefs and desires (rather than objective features of the situation) to predict the player's emotions.

4.5 Personalizing emotion predictions

So far, we have investigated how human observers, and the Inferred Appraisals model, predict emotional reactions for players after observing only a single action in the game. However, the structure of the game means that single actions are highly ambiguous. An observer who knows a specific player might be able to use prior knowledge, from outside the game, to inform inferences about the player's likely values and expectations (Jenkins et al., 2018; Oosterhof & Todorov, 2008). If the Inferred Appraisals model is a good approximation of how human observers reason about players' emotions, it should also be able to predict the emotions observers predict *specific* players will experience.

To mimic prior knowledge of the players, we constructed 20 *SpecificPlayers*, each composed of a unique headshot and brief description. The descriptions included, "Janitor at an elementary school", "Doctor, volunteering in South Africa with *Doctors Without Borders*", and "Investment analyst at a hedge fund" (Figure 4-8D. We hypothesized that even such sparse information would evoke stereotypes that allow human observers to update priors on the players' likely preferences and beliefs.

To test this hypothesis, we asked human observers to rate how much each *Specific-*

Player (e.g. the Eldercare worker, shown in Figure 4-6) actually valued, and valued others believing they valued, *Money*, *AIA*, and *DIA*. For the reputation preferences, observers judged how much the player wanted a reputation for: *"not prioritizing money (people believing that she values other things above maximizing her own personal financial gain)"* (*Money*); *"being considerate (people believing that she does not want to take advantage of her opponent)"* (*AIA*); and *"being competitive (people believing that she does not want to be taken advantage of by her opponent)"* (*DIA*). Observers also rated what the players predicted their opponents would do, given the players' decisions in Split or Steal. After being familiarized with the 'Split or Steal' game, each observer made preference and belief attributes to 8 players, defined by the player's description (a photo and career), the player's choice, and the pot size. Confirming our hypothesis, human observers made consistent and distinct preference and belief attributions to *SpecificPlayers*, which differ from the attributions made to the unspecified *GenericPlayers*. For example, compared to the boxing coach, the eldercare worker is viewed as being more trusting that her opponent will cooperate and less motivated to be perceived as competitive. When these players were shown to have cooperated, the eldercare worker is seen as being less motivated by personal financial gain than the boxing coach.

Overall, the patterns of emotions observers predicted for the *GenericPlayers* were replicated by the emotion predictions made for the *SpecificPlayers* in this experiment. No player is expected to experience more *fury* after winning money than not winning, for example, but how much *fury* observers expect a player to experience differs between players. For example, relative to the *GenericPlayers*, the Eldercare worker to experience more *joy* when she cooperates mutually with her opponent and less *fury* when she gets betrayed. That is, human observers made personalized emotion predictions for each *SpecificPlayer*.

4.5.1 Simulation of the bias induced by personalizing cues

If a model has learned an accurate mapping from inferred appraisals to emotions, then it should be sensitive to variation in the psychological characteristics attributed

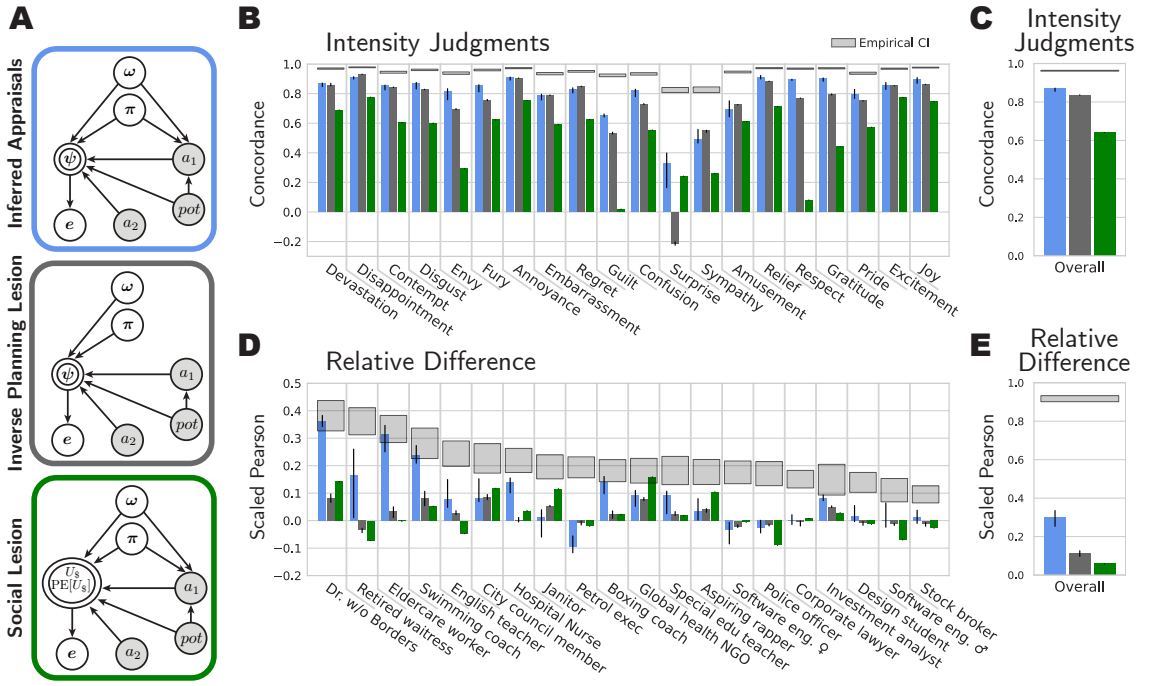

Figure 4-8: **Predicting specific player's emotions.** Human observers made preference and belief attributions to the 20 *SpecificPlayers*, based on a photo, brief description, and decision in the Split or Steal game. Based on what a *SpecificPlayer* was judged to care about and to expect, the models in **(A)** generated predictions of that player's emotion reaction in 24 Split or Steal games (4 outcomes x 8 pots). Bar colors correspond to the models in (A). Grey windows give the 95% bootstrap CI the inter-rater reliability of the emotion predictions. **(B)** Concordance between predictions generated by the models and human observers for every emotion (collapsing across players, outcomes, and pot sizes). **(C)** Overall fit the emotions observers predicted for the 20 *SpecificPlayers*. **(D)** The photos and descriptions of *SpecificPlayers* biased human observers' judgments of the players' motivations, expectations, and emotional reactions. This plot shows how well the models were able to predict the bias in emotion predictions based on observers' judgments of a player's preferences and belief. Players are ordered based on how reliably observers' emotion predictions differed from the emotions predicted for the *GenericPlayers* (grey windows). The model score gives the variance-scaled Pearson correlation. **(E)** Correlation between the relative difference predicted by the models and the relative difference in observers' emotion predictions.

to specific players, which are the bases for inferred appraisals. The key generalization test is therefore whether the Inferred Appraisals model accurately predicts how emotion predictions will differ between players in the same situation, based on observers' ratings of each player's preferences and beliefs.

The expected difference between emotions predicted for a *SpecificPlayer* and the

GenericPlayers is given by $\Delta \boldsymbol{e}_{player} =< \delta^{joy}_{\texttt{CC}},\ \delta^{joy}_{\texttt{CD}}, \ldots >$, where

$$\delta^{joy}_{\texttt{CC}} = \mathrm{E}\left[joy \,|\, \texttt{CC}\,;\ player\right] - \mathrm{E}\left[joy \,|\, \texttt{CC}\,;\ generic\right]$$

In similar fashion, the expected difference predicted by a model, given by $\Delta \hat{\boldsymbol{e}}_{player}$, is the difference between the emotions a model predicted for a *SpecificPlayer* and for the *GenericPlayers*. Since this difference is calculated relative to a model's own prediction of the *GenericPlayers* emotions, a model that fails to fit the absolute expected emotion intensities can still capture how observers' emotion predictions for *SpecificPlayers* change relative to *GenericPlayers*.

The Inferred Appraisals model was able to capture some of the bias in human observers' emotion predictions for *SpecificPlayers*. Across all *SpecificPlayers*, the fit between the predicted difference $\Delta \hat{\boldsymbol{e}}$ of the Inferred Appraisals model and the empirical difference $\Delta \boldsymbol{e}$ was: $ccc = 0.300$ [0.254, 0.335], Pearson $r = 0.314$ [0.266, 0.349]. However, human observers disagreed amongst themselves about how emotion predictions should be personalized for each *SpecificPlayer*. The emotions predicted for some players show more reliable differences from the *GenericPlayers*. We therefore separated the correlations in emotion prediction bias for each *SpecificPlayer* in figure 4-8D. Correlations are scaled by the total variation (see Methods 4.7: Variance-scaled correlation). The Inferred Appraisals model was better able to capture the relative difference in predicted emotions for the *SpecificPlayers* that evoked more reliably different emotion predictions. We hypothesize that the Inferred Appraisals model is mimicking human observers' adjustment of emotion predictions, based on inferred appraisals with personalized values and expectations.

Neither Inverse Planning lesion model nor the Social lesion model were able to generate personalized emotion predictions (Figure 4-8D and E). Despite predicting the expected emotion intensities nearly as well as the Inferred Appraisals model (Figure 4-8C), the Inverse Planning lesion model largely failed to predict how personalizing information biased emotion predictions relative to the generic players: $ccc = 0.112$ [0.097, 0.126], Pearson $r = 0.127$ [0.110, 0.142]. The Social lesion yielded

a still lower correlation: $ccc = 0.058$ [0.058, 0.058], Pearson $r = 0.071$ [0.070, 0.071].

4.6 Conclusion

We propose that human observers predict what emotions others' feel by reasoning about how others' mental contents would cause them to react to hypothetical events. The central premise is that observers infer others' likely appraisals by employing a sophisticated intuitive theory of mind, where the mental contents responsible for someone's behavior also shape that person's emotional reactions. We formalize the theory as a probabilistic generative model and show that observers' judgments about what players in a social dilemma believe, and the utilities they are motivated to optimize, can effectively capture the emotions that players were predicted to experience.

We find that the emotions players were predicted to experience depend on social evaluations beyond the monetary utility that players derive. How much money players were likely to win, and how much they actually won, can partially explain predictions of emotions that covary with the size of the jackpot, including *joy* and *disappointment*, but fail to capture emotions that show little dependence on monetary reward, such as *guilt* and *embarrassment*. Even in the best cases, the predicted emotions are difficult to represent with these monetary predictors. Holding monetary utility constant, for instance, observers predict that the emotions players feel also depend on what their opponents won. Modeling what observers infer about players' social evaluations increases how well a model can capture the 20 emotions observers judged.

The Split or Steal game offers an emotionally evocative yet highly structured paradigm. While many naturalistic paradigms, such as people recounting emotional memories, can offer emotional breadth, these tasks are not well suited for learning quantitative logical structure. By contrast, highly constrained laboratory paradigms, such as lotteries and constructed narratives, can be useful for learning formal structure but typically support only a very limited space of emotion attributions. Rather than proposing a specific formal definition of the intuitive theory of a small number of related emotions, or describing the statistical regularities of a breadth of emotions,

we aimed to capture the logical and structured inference of a range of emotions.

A mental model of other people not only allows observers to reason about people's typical reactions, but also allows observers to infer how a specific person's reactions are likely to differ from other people's. Two players who made identical decisions and ended up in identical situations might still react differently. Observers make systematically different emotion predictions when presented with a photo and verbal description of different players. But what about the player's appearance and career description causes observers to infer that the player is likely to have a different emotional experience than other players? The Inferred Appraisals model proposes that observers' emotion predictions should be understood in terms of causally and logically structured relations between inferred mental contents. The photo and description pairs changed observers' inference about what preferences and beliefs where likely to have motivated the players' choices. The Inferred Appraisals model was able to translate observers' preference and belief attributions to predict the players' emotions, and to some degree, how specific player's emotions were expected to differ. Rather than only inferring that a given player is likely to experience an emotion more intensely or more frequently than another player (a statistical regularity), observers can infer that the player's *fury* has logical and causal structure: Relative to someone else, a player who is inferred to be more competitive (to value his outcome relative to his opponent more, and his absolute monetary reward less) is likely to experience more *fury* when he trusted his opponent and was betrayed, and less *fury* when he and his opponent both made decisions in their own self-interest. The model captured these systematic effects, without being trained on the relationships between inferred preferences and beliefs, and inferred emotions, for the individual players.

Capturing the differences between the emotions predicted for specific players relies on the inverse inference of players' likely preferences and beliefs. Thus, while lesioning inverse planning only slightly reduced the overall fit, the inverse planning lesion heavily impaired the model's capture of how specific players' emotions were expected to differ. Observers use the available cues (in this case, players' faces, career descriptions, and decisions), to infer the players' preferences and beliefs, and reason about

how those latent mental contents would cause the players to react to hypothetical situations. These results argue that an intuitive theory of emotions requires representations with internal structure (Baker et al., 2009, 2017; Davidson, 1963; Skerry & Saxe, 2015), understood in terms of their computational role within a coherent explanatory theory (Carey, 2009; Gopnik & Wellman, 1994).

Despite our model's success, there remain many avenues for future research. First, notwithstanding its richness, 'Split or Steal' affords very limited and simplistic sample of people's emotional experiences. Some emotions are likely to occur only over much longer time-scales (e.g. nostalgia) or only over more existential stakes (e.g. terror). The four types of events create too much covariance between in principle separable inferred appraisals (e.g. 'fairness' and 'respect' are likely to be better representations of people's intuitive theory of preferences (Engelmann & Tomasello, 2019; Starmans et al., 2017), but are too coupled with outcome to be effectively modeled in this paradigm). Even within the world of 'Split or Steal', here we have modeled only emotions that occur in response to the opponent's action, and not the emotions that occur before, in anticipation (e.g. hope, anxiety). Second, the implementation of each of the components of our model could be improved beyond what we have presented here. For example, we use a simple, general prospect function to capture the non-linearity in people's utility derived from the pot size, but future research could replace this with a more accurate empirically derived version of observers' intuitive prospect function. We emphasize that our theoretical commitment is to the computational level relationships rather than the specific implementations. In summary, the integration of appraisal theory, inverse planning, and social utility maximization, in a computational model of people's intuitive theory of mind offers a powerful framework to capture and emulate humans' detailed conceptual knowledge of how others feel.

4.7 Methods

Empirical data collection

Empirical data were collected on Amazon mTurk. Across all experiments, workers were not permitted to participate more than once. Thus, the number of responses contributed by workers is uniform within any dataset, and each dataset was generated by mutually exclusive workers. Financial compensation to workers varied by experiment and targeted 12 USD/hr and was adjusted if experiment duration was misestimated. Workers who began but were unable to finish an experiment (for instance, were unable to view the training video) were paid the full amount when they could be identified. We restricted workers by geolocation to the United States using mTurk credentials and asked workers only begin the experiment if fluent in English. Every experiment included comprehension questions that were used to exclude workers. The comprehension questions varied based on the content of the experiment.

We collected 2 sets of inverse planning attributions (Anonymous Game with *GenericPlayers* and Public Game with *SpecificPlayers*) and 2 sets of emotion attributions (Public Game with *GenericPlayers* and Public Game with *SpecificPlayers*). Planning attributions and emotion attributions were collected from mutually exclusive groups to avoid cueing observers to think about emotion attributions in terms of the planning variables or vice versa. All attribution questions included a photo of player 1 (the target of attribution) that workers were told was taken prior to the contestants revealing their decisions. We emphasized that the photo did not display the players' reactions to the outcome of the game. Workers were told that the jackpots can range from 1 to over 100k USD. For the *Anonymous Game* (one-shot weak PD), Amazon mTurk workers estimated players' three base preferences (ω^{base}_{Money}, ω^{base}_{AIA}, ω^{base}_{DIA}) on continuous scales ranging from the player caring *not at all* to *a great deal* about the feature. Workers also estimated what players believed their opponents were going to choose and how confident the players were (π_{a_2}) on a 6-point Likert-type scale, with 3 ordinal confidence values for each a_2 (e.g. the player thinks her opponent is going

to Steal and is *very confident*, *somewhat confident*, or *not confident*). A `C`/ `D` bias was forced by the even number of response options. Workers completed 8 trials. Each trial included a player's face, the size of the jackpot (24 possible values ranging from 2–20,7365 USD), and the player's decision a_1 to *Split* (`C`) or *Steal* (`D`). Workers were thoroughly informed about the game payoff structure.

For the *SpecificPlayers* we collected emotion attributions as well as preference and belief judgments. Workers were told the payoff structure of the 'Split or Steal' game and shown a video of an on-air negotiation by two real contestants. Preference and belief attributions were collected for every combination of the 20 specific players, 8 pot sizes, and 2 a_1 values, a total of 320 stimuli. The pots are a subset of the 24 collected for the *GenericPlayers*: \$124; \$694; \$1,582; \$5,378; \$12,121; \$27,293; \$61,430 and \$138,238. Workers made inverse planning attributions to 9 stimuli, judging the value of all 7 features (3 base preferences, 3 reputation preferences and 1 belief) for each stimulus. Following a practice trial, workers responded to 8 trials that were balanced across the player's action, the player's gender, and the pot size. In each trial, workers were shown a player's face and occupation, the size of the jackpot, and the player's decision. Individual workers never saw the same player more than once. The initial practice trial was identical for everyone (showing a female 'Customer service assistant' who defected with a 1,090 USD pot), and was excluded from analysis.

Emotion attributions were collected for the same players, pot sizes and a_1 values, with each combination now paired with a specific a_2 value (`C` or `D`), a total of 640 stimuli. Workers made emotion attributions to 9 stimuli, judging the intensity of all 20 emotions for each stimulus. The order of emotions was randomized for each worker. Following a practice trial, workers responded to 8 trials that were balanced across outcome, player gender, and pot size. Individual workers never saw the same player more than once. The initial practice trial was identical for everyone and excluded from analysis. The discarded trailing trial showed the same 'Customer service assistant' won 1,090 USD by defecting (`DC`).

Prior fitting

We integrate out player 1's decision from observers' judgments and smooth the marginal distribution with a Gaussian kernel. For player 1's belief about player 2, we similarly sum over player 1's decision $\mathrm{P}(a_2) = \frac{1}{2}\,\mathrm{P}\,(a_2 \mid a_1{=}\texttt{C}) + \frac{1}{2}\,\mathrm{P}\,(a_2 \mid a_1{=}\texttt{D})$. From observers' empirical attributions of players' preferences and beliefs, we derive 3 sets of priors. In each case, the raw attribution values are smoothed with a Gaussian kernel to yield a kernel density estimate (KDE).

1. The BasePrior was fit using the empirical data collected for the AnonymousGame, which uses the GenericPlayer set of photos (summing out player 1's decision and identity).
2. The GenericPrior was fit using the aggregate of all preference and belief attributions, consisting of the base preferences and belief attributions from the AnonymousGame, and the 7 feature attributions made to all 20 SpecificPlayers in the Split or Steal game (summing out player 1's decision and identity).
3. A set of priors was fit for each SpecificPlayer, conditioned by the player's decision. This produces 40 sets of priors on the 7 planning features. Each set is used to estimate the posterior over appraisal features for that *SpecificPlayer* and decision. For a given *SpecificPlayer* (*player*), inferred appraisals for every pot in the `CC` and `CD` payoffs use the prior $\mathrm{P}\,(\boldsymbol{\omega}, \pi_{a_2} \mid a_1{=}\texttt{C}\,;\, \textit{player})$, with each marginal probability being conditionally independent given a_1 and *player*. For more information, see Appendix B.2.

Probabilistic generative model of inferred appraisals

Additional information about the generative model and calculation of inferred appraisals is given in Appendix B. For each game (a pot-outcome pair), we simulated 3,000 agents. Each agent reacts to the outcome based on its specific beliefs and desires, generating appraisals. The posterior distributions generated based on these variables aim to capture population-level human cognition. The appraisals generated by one simulated agent are not supposed to reflect the inferred appraisals of a single

observer. Rather the aim is for the model posteriors to reflect the combined posteriors of a population of observers.

The model samples from the joint posterior distribution of 19 inferred appraisal variables. These consist of 6 derivatives of each base feature and an updated belief about the opponent. For each base feature, an agent's first-order preference and belief about player 2 give rise to an expected utility (EU) for the choice that the agent makes (a_1). In combination with player 2's choice (a_2), that determines an outcome, which gives rise to an achieved utility (U), a prediction error (PE), a counterfactual calculation on what the achieved utility would have been had the agent made the other choice ($CFa1$), and what utility the agent would have achieved had player 2 made the other choice ($CFa2$), which is what informs prediction error (which are together quantitatively equivalent to the expected and achieved utility we formulate prediction error as the difference between the expected and achieved utility). The agent also incorporates 2nd order preferences on each base feature into its forward planning process, the agent's expected reputation utility (EU^{repu}, or equivalently U^{repu} since there is no update signal supplied by the outcome), which gives rise to the agent's calculation of the reputation utility it would have achieved had it made the other choice ($CFa1^{repu}$). Finally, the latent inferred appraisal space includes the expectation error about the opponent's action, $PE[a_2{=}\texttt{C}]$.

We learn linear functions that translate the inferred appraisals generated by the IA model into predictions of human observers' emotion attributions. Prior to learning a linear transformation, the log-utilities are transformed with a sign-adjusted exponential function. Let $\nu^{-1} = \operatorname{sgn}(U) \cdot (\exp|U| - 1)$, where U is a log-utility. For example, the transformed expected utility for base *Money* is $EU^*[base, Money] = \nu^{-1}\left(\mathrm{E}\,U^{base}_{Money}\right)$, with the star indicating a utility that was exponentiated in this way.

An Inferred Appraisals agent's forward planning procedure estimates the expected utility of each planning feature $\mathrm{E}\,U$ by integrating over the opponent's possible decisions and the agent's belief about how likely the opponent is to make that choice. After the opponent's decision is revealed, the player's achieved subjective utility U follows from the same calculation given absolute certainty about the players' deci-

sions. The contrast between the utility an agent expected to receive, and the utility actually received, gives the agent's reward prediction error. To calculate the difference, utilities values are transformed out of log space while adjusting for the sign. Prediction error is then the linear difference between the utility a player expected verse achieved, $PE^* = \nu^{-1}\Big(U_q(a_1, a_2)\Big) - \nu^{-1}\Big(\mathrm{E}\, U_q(a_1)\Big)$. This formulation of prediction error reflects not just the difference between the absolute reward the agent expected and received, but also how much that agent cares about the feature, such that for the same absolute difference in reward and holding beliefs constant, agents with a stronger preference (e.g. for *DIA*) will experience a prediction error of greater magnitude than their counterparts with a less strong preference.

We calculate counterfactuals that reflect the utility an agent would have derived if player 2 made a different choice (the difference between the achieved utility and the utility that would have been achieved given a different a_2), weighted by how much the agent cares about the feature, and how likely the agent thought the counterfactual a_2 was. This yields larger values for outcomes that player 1 thought were more likely: $CFa2^* = \nu^{-1}\Big(\pi(\neg a_2) \cdot U\,(a_1, \neg a_2)\Big)$

We similarly calculate counterfactuals that reflect the utility an agent would have derived if it had chosen something else, weighted by how much the agent cares about the feature: $CFa1^* = \nu^{-1}\Big(U\,(\neg a_1, a_2)\Big)$

For each base feature, these transformations yield U^*, PE^*, $CFa1^*$ and $CFa2^*$. Reputation features depend on what an agent infers others will infer about its motivations given its decision, but do not depend on what the other player chooses. Since the expected and achieved utilities are identical, each reputation feature yields only U^* and $CFa1^*$.

We transform the appraisals with a prospect theory power function. While it is common for prospect theory power functions to include one to three fit parameters, we use the simplest form, which amounts to a sign-adjusted square root and treats positives and negatives symmetrically: $\mathrm{sgn}(x^*) \cdot |x^*|^\alpha$, where the prospect power term is $\alpha = 1/2$ and x^* could be $PE^*[base, Money]$, $CFa1^*\,[repu, AIA]$, etc. Finally, we include a belief prediction error, $PE[a_2] = \mathrm{P}(a_2{=}\mathtt{C}) - \pi(a_2{=}\mathtt{C})$. The probability of a_2

is certain after the opponent's decision is known.

Maximizing the likelihood of people's emotion attributions under a probabilistic model

We collected empirical emotion attributions by systematically varying both players' decisions and the pot size. For the same combination of independent variables, we sample a vector from the posterior distribution over inferred appraisals. For each combination of independent variables (a_1, a_2, pot), this gives a set of 20-dimensional vectors of people's emotion attributions $\boldsymbol{\mathcal{E}}$ and a set of 19-dimensional inferred appraisal vectors $\boldsymbol{\Psi}$.

We learn a sparse transformation of the joint distribution over inferred appraisal variables into a joint distribution over emotion intensities. The transformation is described by a weights matrix $\boldsymbol{\beta}$ and a diagonal matrix of variances $\boldsymbol{\Sigma}$, which we learn by maximizing the likelihood of observing the empirical emotion vectors, $\boldsymbol{\mathcal{E}}$, under a uniformly weighted multivariate Gaussian mixture. Thus, we fit all the emotions simultaneously and the $\boldsymbol{\beta}$ weights are constrained by the empirical covariance between emotions.

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, b, \boldsymbol{\Sigma}\,;\,\boldsymbol{\mathcal{E}}) &= \prod_t \mathrm{P}(\boldsymbol{e}_t \,|\, \boldsymbol{\beta}, b, \boldsymbol{\Sigma}), \text{ where} \\
\mathrm{P}(\boldsymbol{e}_t \,|\, \boldsymbol{\beta}, b, \boldsymbol{\Sigma}) &= \mathop{\mathrm{E}}_{\mathrm{P}(\boldsymbol{\psi} \,|\, a_1, a_2, pot)} \Big[\mathrm{P}(\boldsymbol{e}_t \,|\, \boldsymbol{\psi}, \boldsymbol{\beta}, b, \boldsymbol{\Sigma}) \Big] \\
&= \mathop{\mathrm{E}}_{\mathrm{P}(\boldsymbol{\psi} \,|\, a_1, a_2, pot)} \Big[\mathcal{N}\Big(\boldsymbol{e}_t\,;\, \mu = \mathrm{logit}^{-1}(\boldsymbol{\beta}{\cdot}\boldsymbol{\psi} + b), \boldsymbol{\Sigma}\Big) \Big] \\
&\approx \frac{1}{N} \sum_i^N \mathcal{N}\Big(\boldsymbol{e}_t\,;\, \mu = \mathrm{logit}^{-1}(\boldsymbol{\beta}{\cdot}\boldsymbol{\psi}_i + b), \boldsymbol{\sigma} I\Big)
\end{aligned}
\tag{4.4}
$$

Where T is the total number of observed empirical emotion vectors $\boldsymbol{e}$, and N the number of posterior samples of inferred appraisal vectors $\boldsymbol{\psi}$ for a given stimulus. We apply a logistic transformation to $\boldsymbol{\beta}{\cdot}\boldsymbol{\psi}$ to accommodate the [0,1] bounds of people's empirical judgments.

Regularization and cross-validation

We fit the weights matrix, $\boldsymbol{\beta}$, via gradient descent with ℓ_1 regularization on every weight. This contrasts with a common approach of seeking a sparse number of total predictors. Rather, each β weight is regularized independently, which allows emotion generation functions to be defined by different sets of inferred appraisal variables. To improve feature selection and the interpretability of $\boldsymbol{\beta}$ weights, the inferred appraisal variables are scaled to have unit standard deviation prior to model fitting.

To learn the scale of the Laplace prior, $\mathrm{P}(\boldsymbol{\beta})$, we performed grid search and K-fold cross-validated on subsets of the *SpecificPlayers*. For every scale of the Laplace prior, the GenericPrior was used to generate inferred appraisal and the $\boldsymbol{\beta}$ weights were fit to the *GenericPlayers* emotion predictions. The SpecificPlayer priors were used to generate inferred appraisals for 3/4 of the *SpecificPlayers*, and the $\boldsymbol{\beta}$ weights were used to transform the *SpecificPlayer* inferred appraisals into emotions. The Laplace scale that provided the best fit to the cross-validation set was then used to predict emotions for the 5 left-out *SpecificPlayers*. This was repeated to generate emotion predictions for all of the *SpecificPlayers*.

Model performance

Concordance correlation

We assessed model performance using Lin's Concordance Correlation Coefficient (*ccc*) (Lin, 1989). Lin's concordance correlation penalizes deviations from the identity line (perfect prediction) making it a more stringent metric than Pearson's Correlation Coefficient. Lin's concordance correlation gives the expected squared perpendicular deviation from a 45 degree line through the origin ($\mathrm{E}[(X-Y)^2]$), which is computed for a sample as,

$$ccc = \frac{2\, s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{4.5}$$

where s_{xy} is the covariance of x and y, s_x^2 is the variance of x and $\bar{x}$ is the mean of x. When no additional centering or scaling can improve the fit of a model, *ccc* is equal to the Pearson's correlation, but the *ccc* is reduced when the model does a poor job of predicting the mean or scale of the empirical data.

Variance-scaled correlation

The variance-scaled correlation scales how well a model predicted observer bias by how different observers expected a *SpecificPlayer*'s emotions to be relative to the *GenericPlayers*. Specifically, $r'_{player} = \text{Corr}\left(\Delta\hat{\boldsymbol{e}}_{player},\ \Delta\boldsymbol{e}_{player}\right) \cdot \left(\sigma_{\Delta\boldsymbol{e}_{player}}/\sigma_{\Delta\boldsymbol{\mathcal{E}}}\right)$ where $\boldsymbol{e}_{player}$ are the attributions made to a specific *player* and $\boldsymbol{\mathcal{E}}$ are the attributions to all of the *SpecificPlayers*. This is a more useful representation of the model performance because the correlation is normalized the variance of the whole data set. If a stimulus is very different from the generic attributions and the model explains it well, the r' will be high. If the stimulus is very different and the model explains it poorly, the r' will be very negative. If the stimulus is not very different from the generic, the correlation is not inflated by the small variance of the empirical deltas.

Chapter 5

Stimulus-computable emotion understanding

> Generalization is thus a cognitive act, not merely a failure of sensory discrimination.
>
> — Roger Shepard (1987), *Toward a Universal Law of Generalization for Psychological Science*

5.1 Introduction

A long-term goal of both psychology and machine learning is to build a model that can match human emotion understanding, from rich naturalistic stimuli of people reacting to emotional contexts. In Chapter 3, we showed that human judgements of what happened, based on an emotional expression, can only be explained by including predictions of emotions in context. In Chapter 4, we built a formal model of predictions of emotions in context, using inverse appraisals as the latent space. In this chapter, we ask whether combining our formal model of emotion prediction in context, with off-the-shelf emotion recognition software, allows us to construct an inferential pipeline to match human observers on tasks that depend on both emotion prediction and expression recognition

We developed two novel tasks to measure how humans reason about others' emo-

tions and expressions in two naturalistic tasks. In the first task (Cue Integration, CI), observers were shown video clips of players on a televised game show, reacting to the outcome of a social strategy game. Based on both a description of the event (the stakes, both players' choices, and the outcomes) and the players' emotional expressions, observers judged how much the player was experiencing each of 20 different emotions. In the second task (Outcome Recovery, OR), observers were told the general structure of the game show and then, based on the video of the players' emotional reactions, asked to guess to what outcome the player was reacting (i.e. what the visible player chose, what the opposing player chose, and the resulting payoffs). Each of these tasks requires observers to combine their expectations for how players will feel, in different game situations, with the players' spontaneous, dynamic emotional expressions, to make judgements similar to the inferences that human observers must regularly make, in the course of social interactions. The overall goal of this project was to measure how well a model can currently capture human performance on each of these tasks, and to identify the most important directions for future improvement. For both the CI and OR tasks, we introduce a stimulus-computable Bayesian model that comprises a perception module, a context module, and an explanation module.

The perception module used off-the-shelf commercial emotion recognition models, Microsoft Azure Emotion Detector and Amazon Rekognition. For the context module, we used a custom model of the causal structure of emotional reactions to events, the Inferred Appraisals model (from Chapter 4), fit to human predictions for events in the same gameshow. These modules feed into the explanation module where Bayesian inference generates the social cognitive reasoning behavior required by a task. The overall results show that the combined model performed moderately well at capturing human observers' performance on both the CI and the OR tasks. For both tasks, combining the perceptual and predictive models is critical; models that used only mappings from perceived expressions to inferred emotions performed substantially worse. The results also illustrate that there remains substantial room for improvement on both the specificity of the perceptual component, and the generality of the prediction component.

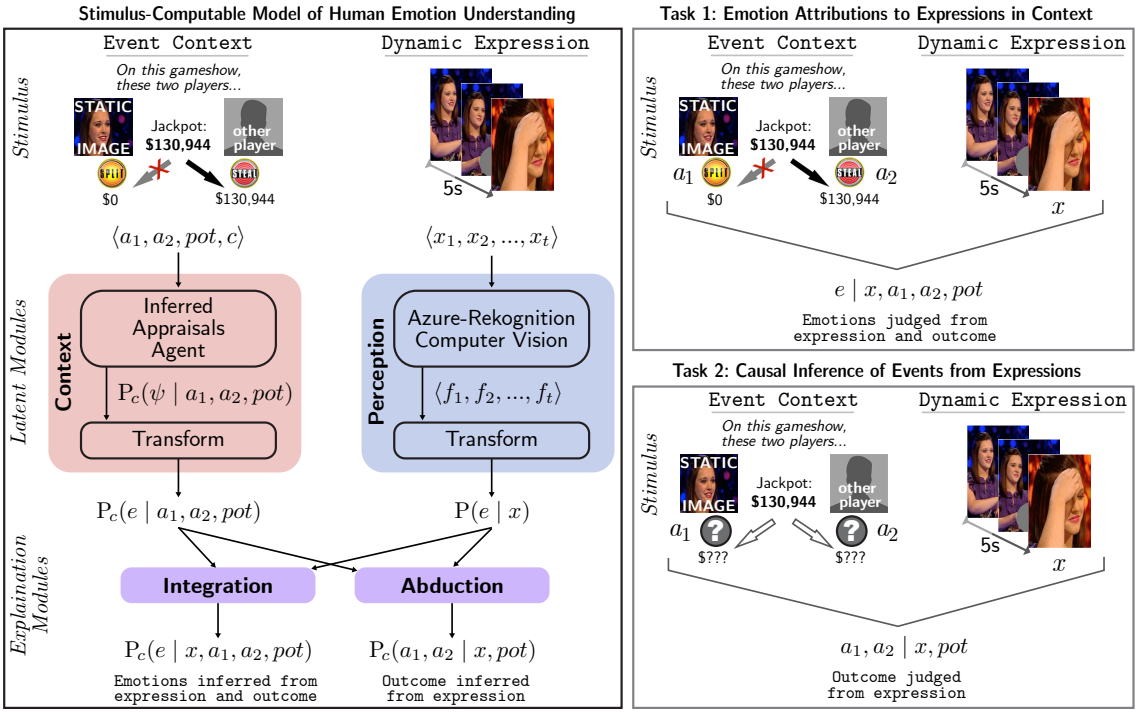

Figure 5-1: Stimulus-computable model of human emotion understanding.

5.1.1 Related work

Advances in deep learning have invigorated efforts to build multimodal models that combine expressive and contextual cues. Multitask models endeavor to capture the flexibility of human emotion understanding by learning a shared embedding that allows partially-observable data to be shared across modalities in service of multiple tasks (Liang et al., 2022). These advances are exciting in that they suggest ways models can perform the variety of social cognitive tasks that humans readily perform (inferring emotions, past events, future actions, personality traits) based on variable and incomplete information from different sources (sometimes facial expressions, sometimes event descriptions, sometimes scene information). These modeling efforts are benefited by datasets that aim to sample human emotion understanding in context (Kosti et al., 2017). Related work aims to learn statistical associations between expressions, scenes, objects, and other people who are visible (Kosti et al., 2019; Parry & Vuong, 2021). These models leverage, often is combination, state-of-the-art deep language models that infer semantic content (Mittal et al., 2020), spatiotem-

poral regularities to infer causal dependencies (Do et al., 2021; Mittal et al., 2021), and graph networks to infer relationships between features (J. Chen et al., 2022; Gao et al., 2021; Park et al., 2020).

In contrast with these models, which learn statistical associations between stimulus features, we frame the problem of emotion understanding as hierarchical reasoning over an intuitive theory of other minds. An intuitive theory is a richly-structured lay theory of ontology, in this case, the generative structure of others' emotions. We build a model of the emotions that observers predict others will experience. Emotion predictions are simulated by inverting a generative model of behavior—how observers explain others' intentional actions in terms of mental contents like wants and expectations. When then model how observers reason about the observed expressions in terms of their contextual knowledge. This theory-based approach treats human emotion understanding as inference over a mental model of psychology, where observers use emotions as latent explanations that connect expressions and events.

5.2 Task 1: Attribution of emotions to expressions in context

In the Cue-Integration (CI) task, human observers attribute emotions to a person based on joint perceptual and contextual cues. Observers viewed a players' dynamic expressions after being given detailed information about the events that the players were reacting to.

5.2.1 Empirical data

In each trial, observers (N=161) were given information about the final round of a GoldenBalls game: the size of the jackpot, the decisions of the two players, and how much money each player won. Observers then saw the dynamic expressions of the focal player. Based on joint knowledge of the events the player experienced and the expressions that the player spontaneously produced, observers judged how much the

player experienced each of 20 different emotions. These data provided a naturalistic target for models of human emotion understanding.

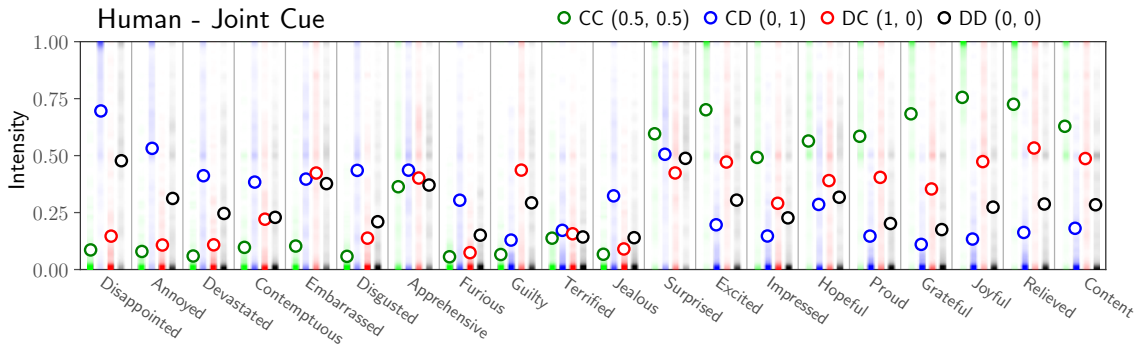

Figure 5-2: **Empirical emotion judgments.** Observers where told what event a player experienced and watch a video of the player's reaction, then rated the player's experience of 20 emotions on continuous scales from "not at all" to "extremely". Points show the expected intensity given the outcome that the player experienced. Shading shows the density of intensity judgments. Legend indicates the proportion of the jackpot that the focal player and the opposing player won. E.g. `CD` (0, 1): the focal player `C`ooperated (a_1=`C`) and won nothing, the opposing player `D`efected (a_2=`D`) and won the entire jackpot.

5.2.2 Model

To explain these human judgements, we built a stimulus-computable Bayesian model (SC-BCI) that comprises a perception module, a context module, and an explanation module. The perception and context modules process the emotion content from a stimulus. The explanation module combines the latent emotion representations output by the latent modules. For the CI task, the explanation model integrates the emotions inferred from expressions and the emotions predicted given event context to infer what emotion intensities are likely given joint access to the cues from both sources.

Perception module: emotions attributed to expressions

To model how human observers interpret the perceptual information conveyed by players' expressions, we used two pre-trained computer vision models of how observers attribute emotions to others' nonverbal expressions. Microsoft Azure Emo-

tion Detector and Amazon Rekognition returned frame-wise emotion ratings for each expression video, which we transformed into the space of emotions judged by human observers. To learn this transformation, an independent group of human observers (N=136) provided emotion judgments based the perceptual information conveyed by players' expressions. The group was told nothing about the GoldenBalls gameshow and judged emotions based solely on the expression videos. We learned a transformation between the joint emotion ratings of Azure and Rekognition and human emotion judgments by iteratively cross-validating on a subset of the expression videos to generate emotion ratings for the left-out players' expressions (see Methods: Perception Module). This results in a 20-dimensional Azure-Rekognition emotion rating for each left-out expression video. Note that "perception" references the input to the module, not the processing of the input. The inference of emotions from expressions involves cognitive processes and is not a strictly perceptual process (Brooks & Freeman, 2018; Brooks et al., 2019).

Context module: emotions predicted from events

To model how human observers interpret event context, we built a generative model of how observers predict others' emotional reactions to situations. The inferred appraisals model predicts emotions from sparse representations of events. In the final round of the GoldenBalls game, the event comprises the actions of two players, the size of the jackpot, and the rules of the gameshow (a public one-shot "weak" prisoner's dilemma). The model simulates human emotion predictions by inverting a nested hierarchical model of observers' intuitive theory of psychology. The model presumes that when observers know what happened to a player, they predict the player's emotions by inferring how the player will cognitively evaluate, or "appraise", the situation on dimensions such as monetary reward, social equity, and public reputation. Based on latent representations of what observers think that players want and expect, the model generates a joint distribution over inferred utilities, prediction errors, and counter-factual judgments.

To learn a transformation between the space of inferred appraisals and human

emotion judgments, we assumed that the joint-distribution over inferred appraisals generated by the model, and the joint distribution over emotions produced by human observers, share a common latent structure. An independent group of human observers (N=164) provided emotion judgments based on descriptions of events that players experienced. This group was told what choices players made, the size of the jackpots, and the rules of the GoldenBalls game, but did not see videos of players' expressions. We learned a transformation between the inferred appraisals and human emotion judgments by iteratively cross-validating on a subset of the players and generated emotion predictions for the left-out players (see Methods: Context Module). This results in a distribution of emotion predictions for each left-out player.

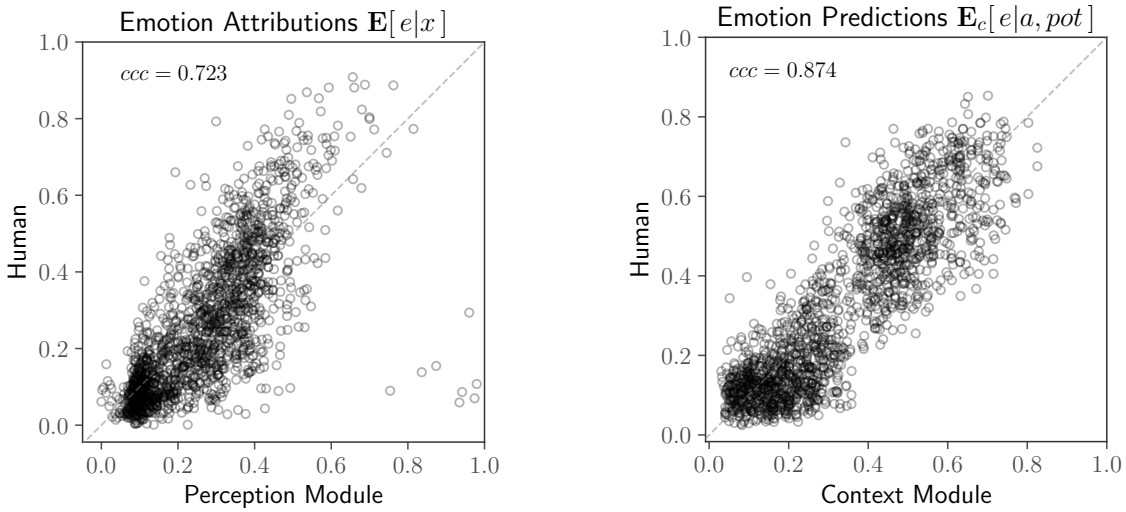

Figure 5-3: **Latent module fits.** The perception and context modules predicted emotions for left-out players. These plots show the expected intensity of the module vs. the empirical data for each emotion of each player. For the perception module, observers attributed emotions to expression videos. For the context module, observers predicted the emotions players would feel based on depictions of the events: the rules of the gameshow (c), the players' actions (a), the size of the jackpot (pot), and a static photo of the player prior to the decisions being revealed.

Explanation module: integration of perceptual and contextual cues

The explanation module is drawn from computational models of human psychology. Ong et al. (2015) proposed that, when human observers have joint access to expressions and event context, they optimally integrate emotion information from perceptual

and contextual cues. In an analogous fashion, we combine the perception module and the context module to build a stimulus-computable Bayesian cue-integration model.

The Bayesian cue-integration model predicts what emotion judgments observers will make given joint access to expressions and event context. The distribution of emotions attributed given joint access to expressions and context, $\mathrm{P}_c(e \mid x, a)$, is proportional to the product of the emotions attributed given expressions, $\mathrm{P}(e \mid x)$, and the emotions predicted given the events, $\mathrm{P}_c(e \mid a)$, and inversely proportional to the prior over the emotions, $\mathrm{P}_c(e)$:

$$\mathrm{P}_c(e \mid x, a) \propto \frac{\mathrm{P}(e \mid x)\ \mathrm{P}_c(e \mid a)}{\mathrm{P}_c(e)}, \quad \text{where} \quad \mathrm{P}_c(e) = \sum_a \mathrm{P}_c(e \mid a)\ \mathrm{P}_c(a) \tag{5.1}$$

Emotions, e, are 20-dimensional vectors, x is the dynamic expression of a player, and a is the outcome determined by the players' decisions. Distributions parameterized by c, as in $\mathrm{P}_c(\cdot)$, indicate that the inference depends on knowledge of the GoldenBalls game, including that it is a public one-shot prisoner's dilemma with the payoff matrix previously described, and the value of the jackpot. We estimate $\mathrm{P}(e \mid x)$ using the perception module, which infers emotions given videos of players' spontaneous expressions. We estimate $\mathrm{P}_c(e \mid a)$ using the context module, which generates emotion predictions given the event context. The single-cue emotion distributions are integrated to simulate $\mathrm{P}_c(e \mid x, a)$, emotion judgments based on joint access to perceptual expression information and conceptual situation information. The hyperprior, $\mathrm{P}_c(a)$, is calculated based on prior knowledge about the statistics of actual gameplay (see Appendix A.2.1).

We also assessed the capacity of the single cue distributions to capture observers' joint-cue emotion judgments. To test if a context-naive expression model can capture observers' behavior, we simulated emotion judgments for each player's expressions using the perception module. The simulated emotion judgments were used to learn a Kernel Density Estimate (KDE) model of the distribution over emotions given each expression video: $\mathrm{P}(e \mid x)$.

We similarly tested how well a model of event context alone can capture observers' joint-cue emotion judgments. The context module generates emotion predictions by inferring players' appraisals using a richly structured model of observers' intuitive theory of mind. Emotion predictions generated by the context module were used to learn a KDE model of the distribution over emotions given the outcome that a player experienced: $\mathrm{P}_c(e \mid a)$.

The Bayesian cue-integration model (SC-BCI), the expression-only model (SC-Expression), and the context-only model (SC-Context) are stimulus-computable, meaning that after training they simulate emotion judgments for novel stimuli without requiring human input. We compare the performance of these models to a human-in-the-loop Bayesian cue-integration model (HITL-BCI), which predicts observers' joint-cue emotion judgments based on empirical emotion judgments of players' expressions and events. Whereas the single-cue models test how well models of perceptual expression information alone and of contextual event information alone capture human emotion understanding, the human-in-the-loop model tests how well Bayesian cue-integration captures human cognition and provides a ceiling for the performance of the stimulus-computable models.

5.2.3 Results

Across all videos, the cue-integration model performed only slightly better than the context-only model and the expression-only model was worst overall. The overall concordance correlation coefficient (*ccc*) between emotion judgments simulated by the Bayesian cue-integration model and human emotion judgments was 0.710. The expression-only model showed a large decrement ($ccc = 0.494$), while the context-only model showed less impairment ($ccc = 0.703$). All three models fell short of the human-in-the-loop cue-integration model's prediction of the empirical joint-cue emotion judgments ($ccc = 0.866$).

However, Bayesian cue-integration was not beneficial in every situation. When players experienced `CD` or `DC` outcomes, the SC-BCI model was less predictive of observers' emotion judgments than the context-only model. In games with these

outcomes, the expression model showed particularly low performance. In contrast with CC, CD, and DC games, where the context-only model performs well relative to the human-in-the-loop model, the context-only model was largely unable to capture how human observers understood the emotions of players in DD games. For these stimuli, the expression-only model showed relatively better performance, and it was beneficial to integrate the expression-only and context-only models.

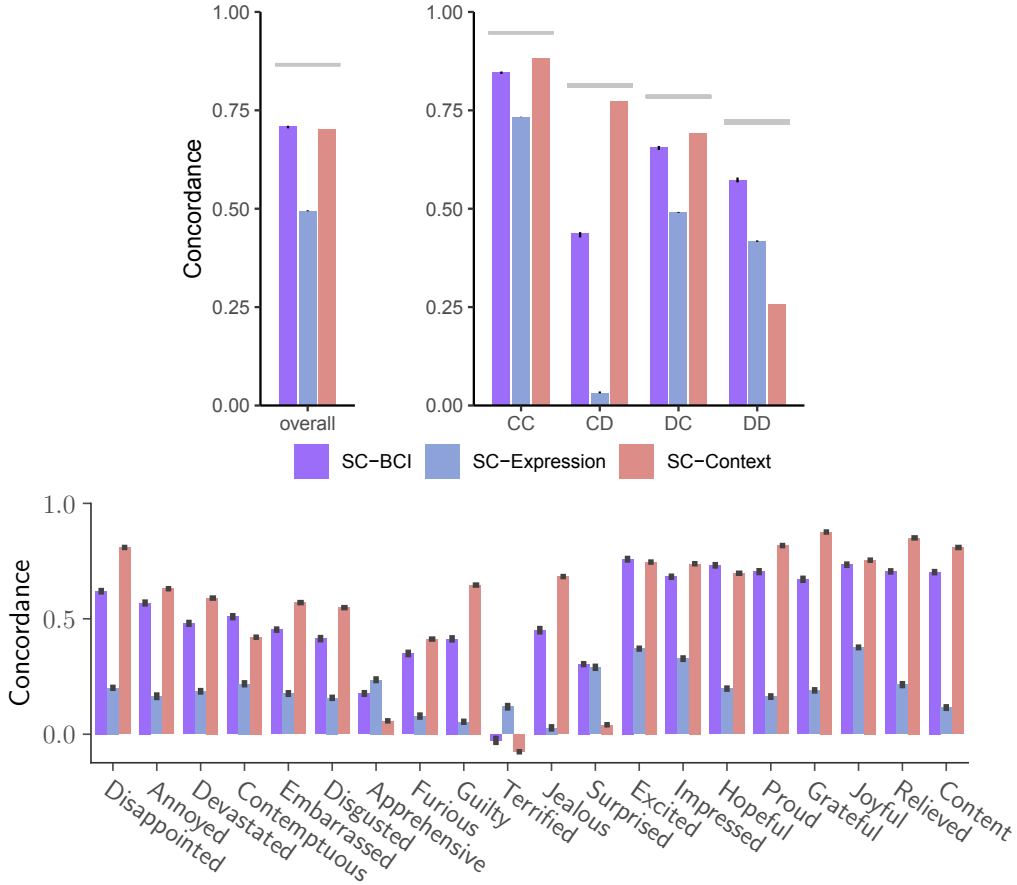

Figure 5-4: **Model fit to human emotion judgments.** The concordance between the inferred emotion intensities and human joint-cue attributions. (top-left) Across all expression videos and emotions. (top-right) Based on which outcome the players experienced. (bottom) Within emotion (across all videos). Shaded windows indicate the 95% bootstrap CI of the HITL-BCI model.

5.2.4 Discussion and Directions

While the SC-BCI model did not regularly outperform both single-cue models, for no stimulus was it the lowest performing model, and it provided the best fit in a number of cases. Emotions inferred by the SC-BCI model were a worse fit to human behavior than emotion inferred by the HITL-BCI model. The SC-BCI model integrates the output of the perception and context modules whereas the HITL-BCI model integrates human emotion judgments, indicating that the stimulus-computable model would be improved if the processing of stimuli was a better match to human cognition. This is particularly evident for the processing of perceptual input.

To measure the extent that more human-like interpretation of expressions would improve the inference of emotions in context, we built a hybrid BCI model by replacing the output of the perception module with the empirical emotion judgments used to train the module. This yields substantially better performance ($ccc = 0.794$ [0.792, 0.795]; Figure 5-5a). Furthermore, how well the perception module predicted human emotion ratings of expressions correlates with how well the SC-BCI model predicted human emotion judgments of the expressions in context. Figure 5-5b shows the concor-

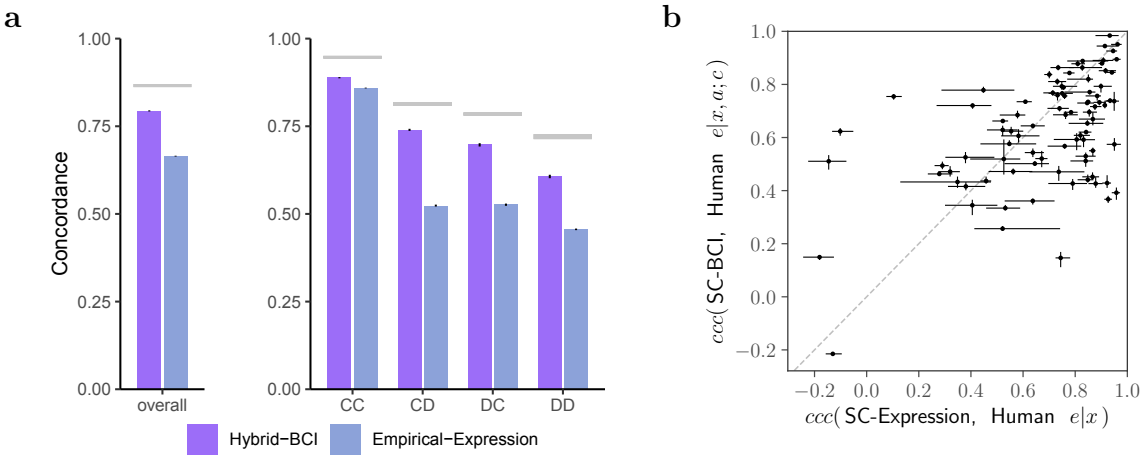

Figure 5-5: **BCI performance with human-level expression processing. (a)** Bayesian Cue-Integration using emotions that human observers attributed to expressions and emotion predictions generated by the context module. Shaded windows indicate the 95% bootstrap CI of the HITL-BCI model. **(b)** Relationship between how well the expression module captured the emotions human observers attributed to expressions versus how well the SC-BCI model captured the emotions humans inferred given joint access to expressions and event context.

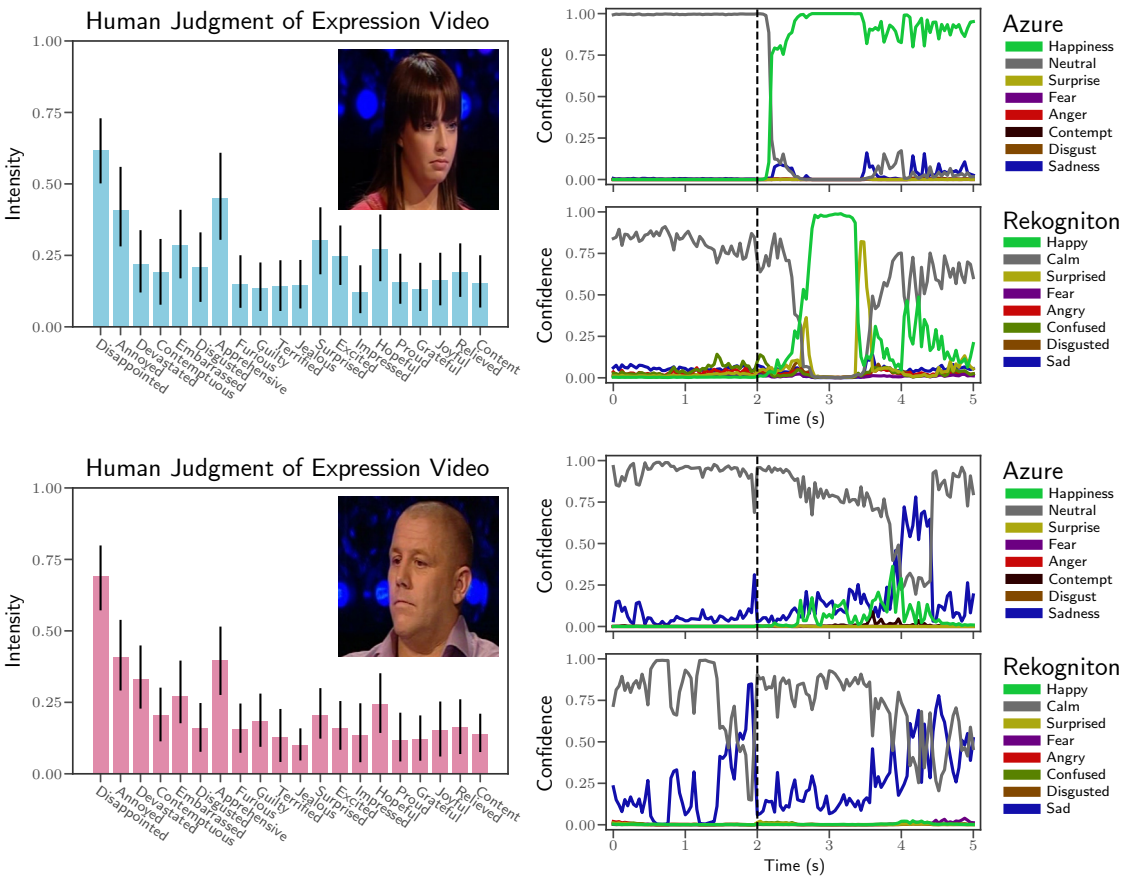

Figure 5-6: **Emotions inferred by humans and computer vision systems.** Human observers attributed similar emotions to these players' expressions, despite the fact that (unbeknownst to the observers), the top player had $19,025 stolen from her (CD) and the bottom player stole $57,518 from his opponent (DC). Observers infer that the CD player is very disappointed, whereas Azure and Rekognition infer that she is happy. For the DC player, Azure and Rekognition exhibit more human-like attributions, inferring that he is sad. These patterns are surprisingly common: human observers frequently infer that players are experiencing very different emotions than the outcomes were expected to elicit, such as being disappointed after winning $57,518. Similarly, the computer vision models frequently deviate from human behavior (e.g. inferring that that the CD player is happy). Where the computer vision models deviate from human behavior, there is no evidence that systems are detecting players' emotions more sensitively than human observers. For example, while we cannot know what emotions the CD player was experiencing, it is exceedingly unlikely that the computer vision models have accurately detected that she is happy about being suckered out of half the jackpot.

dance between the perception module and observers' interpretation of expressions (x-axis) versus the concordance of the SC-BCI model and observers' joint-cue emotion judgments (y-axis). The correlation (Pearson $r = 0.508$) suggests that stimulus-computable prediction of naturalistic human emotion understanding would be sub-

stantially improved by developing models that better match human judgments of expressions. While the development of computer vision models that make human-like emotion judgments of expressions is an active area of research, and many report high levels of success (Ichimura & Kamada, 2022; Khaireddin & Chen, 2021), our results indicate much room for improvement. Figure 5-6 gives examples of the emotions inferred by the computer vision systems and human observers given the same stimuli (i.e. emotion judgments made by observers who saw the expressions without any context information).

5.3 Task 2: Causal inference of antecedent events from expressions

In the second task, human observers see a person's facial expression and infer the unobserved event that elicited that expression. Our model reflects the hypothesis, introduce in Chapter 2 and developed in Chapter 3, that observers predict what emotions someone is likely to experience in each plausible situation and reason about which emotion predictions provide the best causal explanation for someone's expressions. Here, we replace the human-generated perceptual recognition and contextual prediction components in the Bayesian outcome recovery model of Chapter 3 with the stimulus-computable perception and context modules, described above.

The Bayesian outcome recovery model aims to capture the causal structure of how humans reason about expressions. We contend that it is essential to model the emotions that situations are predicted to elicit. An alternative view is that human emotion understanding should be modeled as a function of perceptual patterns of expressive behavior, independent of context. In this view, patterns of expressive behavior are the primary source of information that observers use to understand the emotions, intentions, and situations that caused others' nonverbal reactions (Keltner et al., 2019). Given the availability of powerful tools for building computer vision models, and the challenges of building general models of event context, it is impor-

tant to consider how much of human emotion understanding can be explained by perceptually-available patterns of expressive behavior. As a test of whether human causal reasoning in this task can be explained by perceptual pattern matching, we built a model to classify which outcomes players were reacting to based on the statistical regularities of their expressions alone.

5.3.1 Empirical data

Observers (N=93) viewed players' dynamic expressions and guessed which events the players had experienced. In each trial, observers were told the size of the jackpot, but not what decisions the players made, and then watched the focal player's emotional reaction. Each observer classified the outcome (CC, CD, DC, or DD) for all 88 videos.

Figure 5-7: **Human outcome judgments by stimulus.** Observers guessed which outcome ($a \in \{\texttt{CC}, \texttt{CD}, \texttt{DC}, \texttt{DD}\}$) a player was reacting to based on the player's dynamic expression. Each observer viewed a balanced number of emotional reactions to the four outcome categories, 88 players in total. The bars show the proportion of outcome judgments for each player. Solid bars indicate the proportion of correct outcome judgments and hatches indicate incorrect judgments, with respect to the ground true event. Asterisks indicate expression videos where the 95% binomial CI of the correct judgment is below uniform chance accuracy (0.25).

5.3.2 Model

To explain human causal reasoning about which events the players were reacting to, based on perception of the player's expressions, I built a stimulus-computable model

composed of a perception module, a context module, and an explanation module. The perception module and the context module are identical to Task 1. The explanation module makes abductive inferences about which events provide likely causal explanations of the observed expressions.

Explanation module: abduction of outcomes from expressions

The Bayesian outcome recovery (BOR) model predicts what events observers will infer to have caused players' expressions. This model simulates outcome judgments as abductive inference (Lombrozo, 2012). Observers infer the best explanations for their observations by comparing the emotions players appear to experience against the emotions that hypothetical events are predicted to elicit. The emotions players appear to experience are inferred by the perception module. This yields a distribution of emotions given an expression: $\mathrm{P}(e \mid x)$. The emotions that events are predicted to elicit are generated by the context module. This yields a distribution of emotions given an event: $\mathrm{P}_c(e \mid a)$. For simplicity, we write the two players' actions (a_1 and a_2) as the outcome, a, and include the pot size as a component of the parameterization, c. The explanation module carries out abductive inference over the latent emotion distributions to simulate which events, a, human observers will judged to have caused expression x. This is formalized as a Bayesian belief-updating model, where the probability of event given expression is:

$$\begin{gathered}\mathrm{P}_c(a \mid x) = \int_e \mathrm{P}_c(a \mid e)\, \mathrm{P}(e \mid x)\, \mathrm{d}e \;=\; \mathop{\mathrm{E}}_{\mathrm{P}(e \mid x)} \mathrm{P}_c(a \mid e) = \mathop{\mathrm{E}}_{\mathrm{P}(e \mid x)} \frac{\mathrm{P}_c(e \mid a)\, \mathrm{P}_c(a)}{\mathrm{P}_c(e)} \\ \text{where } \; \mathrm{P}_c(e) = \sum_a \mathrm{P}_c(e \mid a)\, \mathrm{P}_c(a)\end{gathered} \tag{5.2}$$

Thus, the model infers the posterior distribution, $\mathrm{P}_c(a \mid x)$, by reasoning about the likelihood that $\mathrm{P}(e \mid x)$ was sampled from the conditional distribution $\mathrm{P}_c(e \mid a)$. As was the case with the SC-BCI model, the hyperprior, $\mathrm{P}_c(a)$, is calculated based on prior knowledge about the statistics of actual gameplay (see Appendix A.2.1).

We compare the performance of the SC-BOR model to human-in-the-loop Bayesian

outcome recovery model (HITL-BOR), which predicts observers' outcome judgments based on empirical emotion judgments of players' expressions and emotion predictions based on the events. This human-in-the-loop outcome recovery model provides an estimate of how well Bayesian outcome recovery can capture human causal reasoning about the event antecedents of expressions and provides an estimate of the upper limit that can be achieved by stimulus-computable models.

5.3.3 Results

The Bayesian abductive inference module predicts which events observers infer a players experienced, based on videos of the players' spontaneous emotional expressions. For a given expression video x, the model infers the probability that human observers will judge that the player is reacting to outcome a. Model performance is calculated by comparing the empirical and simulated $\mathrm{P}_c(a \mid x)$ for every combination of x and a. For instance, we compare the proportion of human observers who judged that player i was reacting to outcome `CC` with the model inference of $\mathrm{P}_c(a{=}\texttt{CC} \mid x{=}i)$ (the posterior probability that observers would judge `CC` given the dynamic expression of player i). Across all outcomes and players, the outcome judgments simulated by the model and the outcome judgements of human observers showed a concordance correlation of 0.523 [0.503, 0.541].

5.3.4 Perceptual outcome classification

The expression model classifies which outcome a player experienced based on the expressions that players actually made. This simulates the outcomes human observers would infer if they had an accurate understanding of what expressions players produce. A Support Vector Machine (SVM) was trained on the Microsoft Azure and Amazon Rekognition emotion ratings of players' expressions to predict which outcome the players were reacting to. Since the SVM does not require the expression information to be in the 20-dimensional empirical emotion space, we used the summary statistics of the Azure and Rekognition time series. Thus, the SVM bypasses

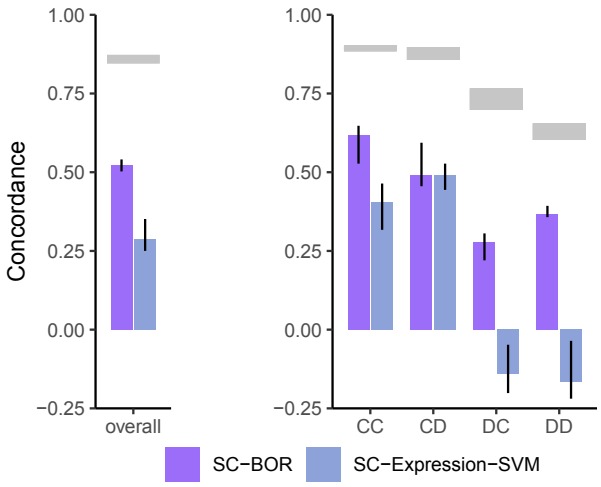

Figure 5-8: **Fit to human outcome judgments.** Concordance between the inferred probability of an outcome judgment and the proportion of observers who guessed that outcome. (left) Across all expression videos. (right) Based on the ground truth of which outcome the player experienced. Shaded windows indicate the 95% bootstrap CI of the HITL-BOR model.

the transform step of the perception module.

Results

Perceptual classification proved to be a poor model of human behavior. Across all outcomes and players, the outcome judgments simulated by the expression model showed a concordance correlation of 0.289 [0.250, 0.352] with the outcome judgements of human observers. Divergence between the expression model human behavior was especially pronounced for the players reacting to `DC` and `DD` games (*ccc* = -0.142 [-0.202, -0.048] and -0.166 [-0.219, -0.035], respectively).

5.3.5 Discussion and Directions

The OR task is a useful test of model performance because it involves social cognition about expressions, rather than only direct rating of expressions. Outcome judgments simulated by the SC-BOR model were a much better fit to human behavior than outcome judgments simulated by the expression model (SC-Expression-SVM). Thus, incorporating contextual information and causal structure yields more human-like

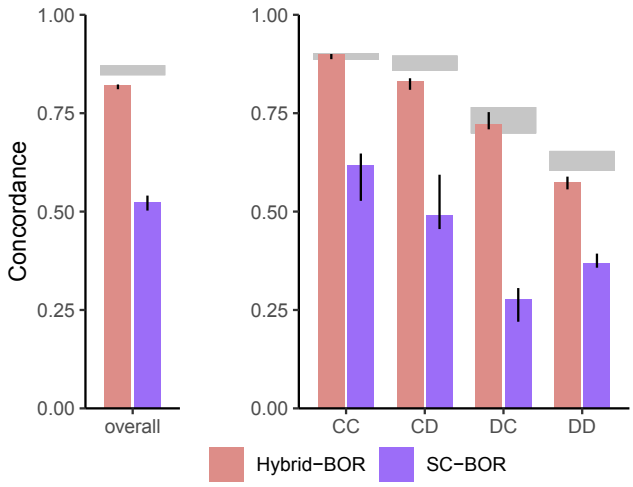

Figure 5-9: **BOR performance with human-level expression processing.** Bayesian Outcome Recovery using emotions that human observers attributed to expressions and emotion predictions generated by the context module. Shaded windows indicate the 95% bootstrap CI of the HITL-BOR model.

expression judgments. Compared to the complicated, recursive, and unconstrained social cognition that people naturalistically employ, the OR task is extremely simple. It is likely that in most tasks of even moderate complexity, isolated expressions will be insufficiently informative for stimulus-computable models to emulate human social cognition.

Similar to the results in the CI task, performance of the SC-BOR model was substantially lower than the human-in-the-loop comparison model, HITL-BOR. To measure the extent that more human-like interpretation of expressions would improve the inference of outcomes, we built a hybrid BOR model by replacing the output of the perception module with the empirical emotion judgments used to train the module. This improved the performance to nearly the level of the HITL-BOR comparison model: Hybrid-BOR $ccc = 0.820\,[0.811, 0.823]$; HITL-BOR $ccc = 0.863\,[0.849, 0.869]$. Thus, the difference between the Perception Module and human observers accounted for most of the error of the stimulus-computable model relative to the human-in-the-loop model.

We now consider accuracy with respect to ground truth. Chapter 3 showed that observers make cognitive model-based errors in their causal reasoning about these

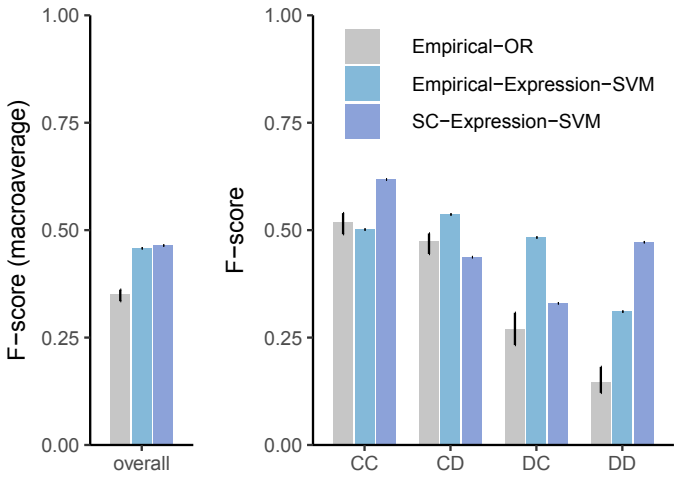

Figure 5-10: **Accuracy with respect to the ground true events.** Empirical-OR: The F1-scores of human observers' outcome judgments. Error bars give the 95% bootstrap CI on the performance of individual observers. Empirical-Expression-SVM: trained to classify the true outcome based on empirical emotion judgments of expression videos. SC-Expression-SVM: trained to classify the true outcome based on summary statistics of the computer vision systems ratings of expression videos.

players' spontaneous emotional expressions. One possibility is that expressive behavior in simply uninformative. It is also possible that expressive behavior encodes relevant information and humans are either insensitive to, or do not accurately interpret, the relevant perceptual cues. Indeed, prior work has shown that observers fail to utilize statistical regularities in expression production. Using static photos, Aviezer et al. (2015) found that observers' ratings of emotional valance were insensitive to objective differences between expressions of tennis players who won or lost a point. Thus, while supervised learning to classify veridical events from players' expressions yields a poor fit to human behavior (Figure 5-8), computer vision systems may be able to recover the true outcomes better than humans.

We repeated the perceptual outcome classification analysis from above (Section 5.3.4), but this time the SC-Expression-SVM model was trained to recover the ground-true outcome rather than human outcome judgments. We trained another SVM, the Empirical-Expression-SVM model, to recover the ground-true outcome from human emotion judgments of the expressions (the same empirical data used to train the Perception Module). The macroaveraged F1-scores of the ground-truth super-

vised SVM models were comparable: 0.464 and 0.458 for SC-Expression-SVM and Empirical-Expression-SVM, respectively (Figure 5-10). Both models outperformed human observers in classifying which events elicited players' spontaneous expressions (macroaveraged F1-score and 95% CI of human observers: 0.350; Empirical-OR in Figure 5-10).

The results indicate that these computer vision systems are not more sensitive to diagnostic expressive signals than human observers, since both sets of emotion judgments support similar classification performance. Additionally, the results indicate that there are statistical regularities in spontaneous expressions that differentiate which antecedent events players experienced, but that individual observers fail to use this diagnostic information in their causal reasoning. Together this suggests that human observers could learn to interpret the diagnostic information to perform better on the OR task, but that the computer vision systems should not be expected to exceed human-level performance.

5.4 General Discussion

Our goals are to build a stimulus-computable model of emotion understanding, to highlight the areas currently preventing computational models from achieving human-like behavior, and to suggest directions for improvement. Emotion understanding is likely to be an intermediate step in many aspects of social cognition. Others' emotions provide vital information about what they care about, what they are likely to do in the future, what events have transpired previously, and how we should plan our own actions (de Melo et al., 2014; Houlihan et al., 2022; Saxe & Houlihan, 2017; Wu et al., 2021).

It follows that capturing how people *use* emotion information is likely to be a meaningful test of whether models have captured the cognitive structure of human emotion understanding. In this spirit, we examined two naturalistic tasks that humans regularly perform: inferring people's emotions given joint access to their expressions and the events they are reacting to, and what events people are reacting to

given their expressions. These tasks require observers to combine and reason about emotion information. In the current work, we model human cognition by using a perceptual front end to process expressions, a generative model of inferred appraisal to process event context, and hierarchical Bayesian models to computationally reason over latent emotion information in order to find likely explanations for what was observed.

Our work argues that these components are crucial for any model of emotion understanding that aims to simulate human behavior in naturalistically rich social situations. Within limited domains, it is easy for individual components to perform deceptively well. Current computer vision models are already reporting high levels of success at predicting human annotator emotion labels for static facial expressions scraped from search engines (Ichimura & Kamada, 2022; Khaireddin & Chen, 2021). By contrast, when the contextual information is much richer than the expression information, context alone can appear to be a better predictor of emotion understanding (Goel et al., 2022). The CI task illustrates that expressions can appear more informative in certain cases (e.g. for the expressions of players in `DD` games) while context appears more informative in other cases (`CC`, `CD`, and `DC` games). Building artificial systems that capture human emotion understanding will invariably require modeling both expressions and event context, as well their interactions. The challenges involved in advancing the individual components may be locally distinct, but are fundamentally interconnected (Saxe & Houlihan, 2017).

In the present work, the stimulus-computable model would be most improved by a perception module that more closely matched human judgments of expressions. Several factors would likely lead to better approximation of human emotion attributions to expressions. Temporal integration of facial and bodily expressions is an active area of research for computer vision, but remains a challenge (Do et al., 2021; Jin et al., 2020). Humans show complex processing of spatiotemporal statistics of expressions (Goldenberg et al., 2022; Jack et al., 2014; Sowden et al., 2021). The computer vision models used in the current work processes video frames independently. While the temporal integration strategy we used is common, it is likely a poor approximation of

human cognition. Similarly, body posture can dramatically influence human emotion judgments (Aviezer et al., 2012b; Israelashvili et al., 2019; Lecker et al., 2020). It was not uncommon for players in our stimuli to bring their hands up to their mouths and faces. Seeing someone bury her face in her hands is a meaningful gesture to human observers, which underscores the need for robust, multimodal computer vision systems that can not only tolerate occlusions, but also interpret them. There is also evidence that the space of emotions that observers reliably attribute to expressions is larger than the low dimensional annotation sets typically used to supervise the training of computer vision models (Cowen & Keltner, 2020). In the present work, we learned a mapping between the eight-emotion time series output by the computer vision models and observers' intensity judgments of twenty fine-grained emotions, but future computer vision models should aim to capture a much richer space of the emotions that humans interpret in others' expressions.

The emotions reported by Azure and Rekognition diverged strikingly and qualitatively from the emotion judgments of human observers. One possible explanation is that our current stimuli are highly dissimilar from the distribution of images in the training set. Data used to train computer visions models, which is often scraped from internet search engines, can reflect potentially extreme selection biases (Peña et al., 2020). Computer vision systems may be better able to generalize to spontaneous expressions if training data more accurately reflected the distribution of naturalistic expressions (Le Mau, 2019; McDuff et al., 2019) and if the models explicitly aimed to learn observers' knowledge of the naturalistic variability in expression production (Abdíc et al., 2016; Anzellotti et al., 2021; Barrett et al., 2019; Lei & Gratch, 2019; Martin et al., 2017). The especially poor fit to certain players' expressions is likely due in part to the players' demographics. Even relative to human observers, computer vision systems are prone to making emotion inference that are biased by race, gender, and age (Bryant & Howard, 2019; Kyriakou et al., 2020; Xu et al., 2020). While there is ongoing work to curate better training data (Y. Chen et al., 2021; Prabhu & Birhane, 2020), mitigating these issues is a pressing ethical concern that will likely require critical examination of the current approaches and practices (Birhane, 2021;

Ong, 2021; Scheuerman et al., 2021).

The context module faces different challenges than the perception module. The context module provided an excellent match to how observers predict the emotions that players will experience in these tasks. In contrast to the computer vision models, which aim to learn general mappings between expressions and emotions, the Inferred Appraisals Agent model was narrowly designed to capture inferred appraisals in one-shot, public Prisoner's Dilemma games. Within the training and test sets of events, the events and objective rewards were highly constrained: there was a limited set of possible actions, and the objective rewards were known. A more general context module would need to operate over complicated and abstract multimodal cues. For instance, observers readily predict emotions from descriptions and depictions of situations (Le Mau et al., 2021; Skerry & Saxe, 2014, 2015). Recent advances in deep learning have permitted language models to capture an impressive amount of relevant social information from relative unconstrained stimuli like these (Bhagavatula et al., 2019; Bosselut et al., 2019; Park et al., 2020). Conversely, there's evidence that the learned embeddings fail to capture human-like mental state representations (Shu et al., 2021; Stojnic et al., 2022). In our view, theory-based symbolic reasoning is an indispensable part of human-like emotion understand models and may also be learnable with the right inductive biases. The path towards more human-like models of context-based emotion understanding necessitates combining symbolic generative models with deep learning and reinforcement learning frameworks (Le et al., 2021; Ong et al., 2021; Tsividis et al., 2021).

Clearly, it will take a great deal of innovation to develop stimulus-computable models that capture human emotion attributions to expressions and human emotion predictions from situations. But on its own, this would not be sufficient to explain human emotion understanding in the current tasks. Even the HITL models, which directly used observers' emotion attributions to expressions and emotion predictions from events, did not perfectly capture the target data. In the present work, we approximated the latent emotion distributions using the explicit emotion judgments of observers given partial information. This framework presumes more independence

between the perception module and the context module than human behavior implies.

Our current data indicate where modeling rich latent emotion representations are likely necessary. We have previously argued that observers intuitively understand some of the real-world variability in expression production (Anzellotti et al., 2021). While observers might not be able to accurately introspect on their knowledge, there is evidence that observers' latent probabilistic representations are a significant factor in how they combine and reason over emotion cues. In the CI task, the explanation module uses the population-level judgments to approximate the latent emotion distributions of individual observers. For some combinations of expressions and context, Bayesian integration of the single-cue emotion judgments clearly fails to capture how human observers interpret expression cues in the light of event context. An example is shown in Figure 5-11.

Similarly, in order to capture human causal reasoning in the OR task, a stimulus-computable model would need predict how contextually-informed hypotheses shape

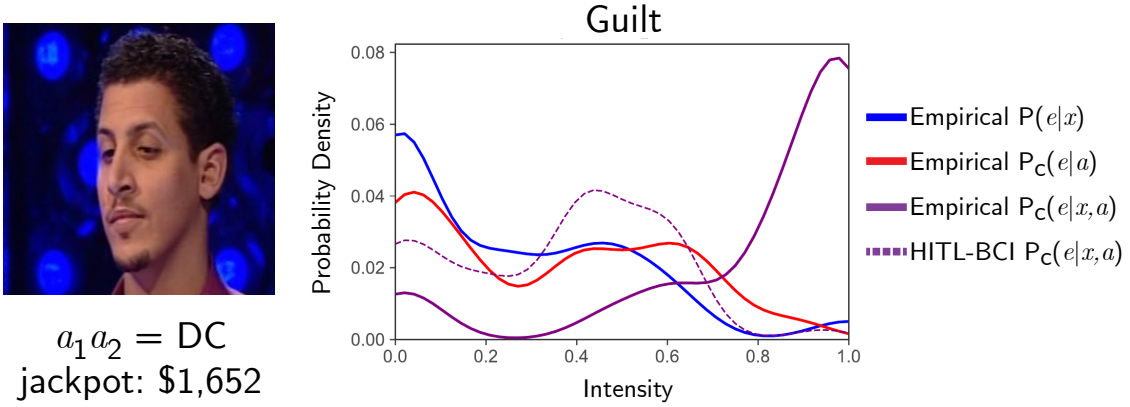

Figure 5-11: **Limits of population-level Bayesian cue-integration.** The blue line shows the probability density of human observers' attributions of *guilt* based on the perception of the player's expression. Given his expression alone, most observers inferred this player was not experiencing much, if any, *guilt*. Similarly, when observers knew what happened (he `Defected`, his opponent `Cooperated`, the jackpot was $1,652) and saw this photo of him (taken before the players revealed their choices), they did not predict that he would experience much, if any, *guilt* (red line). Bayesian integration of these distributions predicts that observers who saw his dynamic expression and knew what event he was reacting to, would infer that he was experiencing a moderate amount of guilt (HITL-BCI, dashed purple line). However, observers with joint access to his dynamic expression and the event context inferred that he felt *extremely* guilty (solid purple line).

the interpretation of expression cues. Future stimulus computable models will need to not only capture the mean emotion judgments, or even the distribution of a population's judgments, but also the latent probabilistic representations that observers use to reason about ambiguous information. In our view, this requires generative models that computationally recapitulate the human intuitive theory of emotion.

5.5 Conclusion

Human emotion understanding is a theory-based solution to the problem of explaining ambiguous observations with incomplete knowledge. From sparse, uncertain data, observers make rich inferences and predictions about other people and the world. We have argued that observers solve these ill-posed inverse problems using a rich, causally-structured mental model. The structure of this mental model is echoed in the emotions observers attributed to players' expressions. Observers who knew what space of events players could be reacting to made different emotion inferences than observers who knew nothing about the provenance of the expressions. Furthermore, this shift in interpretation was important for capturing how observers causally reasoned about the events that players experienced. We take this as evidence that hypotheses about mental states constrain the search for likely explanations of the available data. In this case, the highly constrained event structure leads to different interpretations of facial and bodily musculature. When the situation is less constrained, observers almost certainly use expressions to guide the search for likely world states. Expressions that seem incongruent with one set of contextual hypotheses likely impel observers to search for alternative world states that can better explain the observed reaction. Therefore, while treating the perception and context modules as independent and parallel is a secondary concern for the current work, stimulus-computable models of human emotion understanding will need to address how conceptual and perceptual information mutually and recursively constrain the interpretation of expressions and events.

It is possible that end-to-end multimodal models that learn joint embedding of

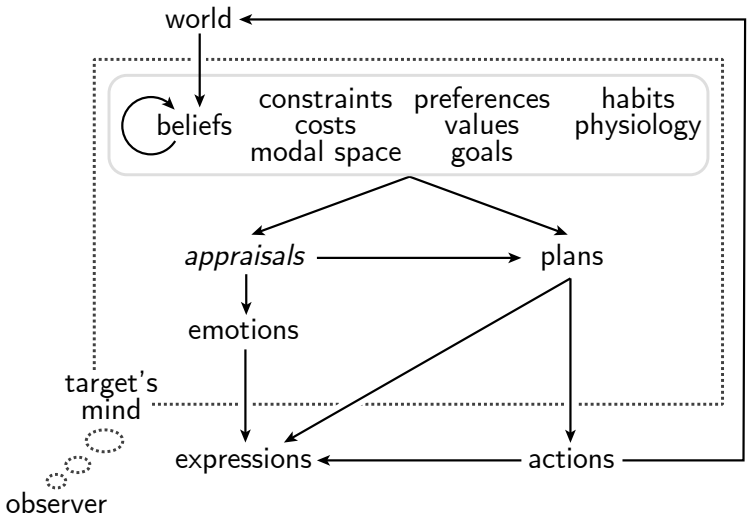

Figure 5-12: **Schematic of the intuitive theory of emotions.** A causally-structured framework for developing generative models of social cognition (for details, see Chapters 2 and 4). Formal models of intuitive psychology have traditionally relied on hand-coding explicit symbolic representations (e.g. agents, goals, efficiency, appraisals). While such structure is likely required for models to match the generality, sophistication, and adaptability of human cognition, this structure may be learnable under the right inductive biases.

expressions and context will be able to capture how expressive behavior and contextual information mutually constrain emotion understanding. However, in our view, statistical associations between expressive behavior and contextual cues are a poor approximation of how humans reason about other minds. Capturing the fine-grained logical and causal reasoning that typifies social cognition will require the use of probabilistic generative models of human's intuitive theory of mind (Lake et al., 2017). While generative models, like the Inferred Appraisals Agent model, have historically relied on extensive hand-coding, advances in program induction indicate that probabilistic programs of this type can be learned (Lake et al., 2015).

In our view, building human-like interpretation of expressions will require incorporating generative models of observers' intuitive theory of emotions (Figure 5-12). Rather than treating expression interpretation as a problem of perceptual pattern matching, it should be treated as causal reasoning where observers invert an intuitive theory of expression production to infer mental contents (appraisals, beliefs, desires, motivations, costs). These latent theory of mind variables are the grammar of social cognition. Models that aim to capture how humans combine, reason over, and act

on emotion information, will invariably need to model computations in the space of these latent mental contents.

5.6 Methods

Stimuli

The stimuli were generated from archival footage of a televised British gameshow called *GoldenBalls*. Every episode of GoldenBalls culminates with two contestants playing a dramatic one-shot instantiation of the Prisoner's Dilemma. Each player is given a choice to "Split" or "Steal" a jackpot (in standard Prisoner's Dilemma notation, to "`C`ooperate" or "`D`efect", respectively). If both decide to "`C`ooperate", they each receive half of the jackpot. If one player instead chooses "`D`efect", that player wins the entire jackpot and the other player who chose "`C`ooperate" leaves with nothing. If both players choose "`D`efect", both get nothing[1]. Players negotiate with each other in front of a live audience in an attempt to convince the other to make the financially disadvantageous choice to `C`ooperate. Each player makes a decision in private, then the two players simultaneously reveal their choices, all while being filmed. The game is emotionally evocative by design. When the choices are revealed, players discover whether they have won or lost real and often substantial sums of money; and whether they have successfully cooperated, successfully duped, been duped by, or failed to dupe the other player. The TV cameras capture their spontaneous, unscripted expressions.

The stimuli for both novel tasks were depictions of 88 individual players' experiences in the Split of Steal game. The four outcomes were represented equally in the stimuli (N=22 for each outcome), reflecting the true distribution of play (across all 287 broadcast episodes, players were 53% likely to cooperate (van den Assem et al., 2012), and the decisions of a player dyad were statistically independent of each other

[1]Rapoport (1988) defines this payoff structure as a Weak Prisoner's Dilemma because the `CD` payoff confers the same monetary reward ($0) to player 1 as the `DD` payoff. Thus, with respect to a player's first-person financial payout, defecting is never harmful, but is only conditionally beneficial.

(Burton-Chellew & West, 2012)). We decomposed these experiences into elements of *Expression* and *Context*.

To convey the focal player's expression, we created a 5-second video of each player, by splicing together footage from the moments surrounding the climactic reveal. Each silent video shows a single player's expressions; the first 2-seconds show the player before the decisions are revealed and the final 3-seconds show the player's reaction to learning the game outcome. The players' decisions were masked so that observers could not see what outcome the players were reacting to. The focal player (the player visible in a video) is always coded as player 1. These 88 videos were used as stimuli in both Task 1, attributing emotions to expressions in context, and Task 2, inferring specific contexts from expressions.

To convey the focal player's outcome, a static graphic depicted the pot size, the focal player's choice, the other player's choice, and the outcome for each player. Videos were presented with the ground-truth outcome display. This graphic was presented along with the corresponding video to human observers in Task 1 only.

Perception Module

The module simulates emotions attributed based on the perception of a player's spontaneous expression, in the absence of contextual information: $\mathrm{P}(\boldsymbol{e} \,|\, x)$. From a 5-second expression video, x, still frames were extracted. For every video frame $\langle x_1, \cdots, x_{t=126} \rangle$, the computer vision component returns an interpretation of the pixel-level visual statistics $\langle f_1, \cdots, f_{t=126} \rangle$. Microsoft Azure Emotion Detector and Amazon Rekognition each return an 8-dimensional simplex vector reflecting the model's confidence of detecting 8 model-specific emotions. For each video, the time series of each f feature is converted into a summary statistic: the mean of the anticipation expressions (using frames from the first 2-seconds of a video), and the mean and the max of the reaction expressions (using frames from the last 3-seconds of a video). This yields a 48-dimensional summary feature vector for each 5-second expression video.

We learn a linear transformation between the summary feature vectors and the empirical 20-dimensional emotion intensity judgments using Ridge regression. The

data for 4 of the 88 expression videos (one from each outcome category) were held out for testing. We fit the transformation parameters with L_2-regularized linear regression. The regularization hyperparameter was fit by K-fold cross-validating on the training data using grid search. The process was iterated over data partitions to simulate out-of-sample emotion attributions for every expression video.

Context Module

The module simulates emotion predictions based on the event context, in the absence of perceptual information about a player's expressive reactions to the event: $\mathrm{P}_c(\boldsymbol{e} \mid a_1, a_2, pot)$. We generated emotion predictions following the method described in Chapter 4. Given the actions of a player dyad and the size of the jackpot, the model generated a joint distribution over inferred appraisals: $\mathrm{P}_c(\boldsymbol{\psi} \mid a_1, a_2, pot)$.

Distributions parameterized by c, as in $\mathrm{P}_c(\cdot)$, indicate that the inference depends on knowledge of the GoldenBalls game, which corresponds with the "Public Game" model in Chapter 4. Inferred appraisals were generated using empirically-derived prior over players' preferences and belief (the "GenericPrior", Chapter 4)[2].

We learn a sparse transformation between the joint distribution over inferred appraisals and a joint distribution over empirical emotion intensity judgments (an empirical KDE):

$$\boldsymbol{\beta}, \boldsymbol{b} = \arg\max_{\boldsymbol{\beta}, \boldsymbol{b}} \prod_{m}^{M} \prod_{n}^{N} \mathrm{P}(\boldsymbol{e}_{m,n} \mid \boldsymbol{\psi}_m \,;\, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\Sigma})\, \mathrm{P}(\boldsymbol{\beta}) \tag{5.3}$$

where M is the number of stimuli and m specifies the stimulus. A stimulus is defined by the event context, $\langle a_1, a_2, pot \rangle$. N is the number of empirical emotion judgments of stimulus m. The likelihood function is described in Methods 4.7: Maximizing the likelihood of people's emotion attributions. Note that the 20 emotion labels for which the Inferred Appraisals model predicts intensities are different than the 20 emotion

[2]In the present chapter, the Inferred Appraisal model generates emotion predictions for different veridically stimuli, where the players' actions, jackpots, and static photos of players' faces reflect the ground-true pairings in the GoldenBalls gameshow. However, the GenericPrior used here and described in Chapter 4, $\mathrm{P}_c(\boldsymbol{\omega}^{base}, \boldsymbol{\omega}^{repu}, \pi_{a_2})$, was estimated by marginalizing over combinations of actions, pot sizes, and player identities that did not actually occur on the GoldenBalls gameshow.

labels in Chapter 4.

The data for 8 stimuli (balanced by outcome category) were held out for testing. The maximum *a posteriori* (MAP) estimate of the transformation parameters was learned on the remaining data. We fit the weights matrix, $\boldsymbol{\beta}$, via gradient descent with L_1 regularization on every weight. The scale of the Laplace prior, $\mathrm{P}(\boldsymbol{\beta})$, was learned through K-fold cross-validation on the training data using Bayesian optimization (Snoek et al., 2012). The MAP estimate of the transformation parameters was then used to generate emotion predictions for the held-out stimuli. The process was iterated over data partitions to simulate out-of-sample emotion predictions for every event description.

Models

Model	Expression Processing	Context Processing
Cue-Integration (CI) Task		
HITL-BCI	Humans	Humans
SC-BCI	Perception Module	Context Module
SC-Expression	Perception Module	✗
SC-Context	✗	Context Module
Hybrid-BCI	Humans	Context Module
Empirical-Expression	Humans	✗
Outcome Recovery (OR) Task		
HITL-BOR	Humans†	Humans
SC-BOR	Perception Module	Context Module
SC-Expression-SVM	Perception Module‡	✗
Hybrid-BOR	Humans	Context Module

Table 5.1: **Key for models.** Stimulus-computable (SC), human-in-the-loop (HITL), Bayesian cue-integration (BCI), Bayesian outcome recovery (BOR). For the HITL-BOR model, † indicates that observers knew the rules of the GoldenBalls gameshow when judging the emotional content of players' expressions. For the SC-Expression-SVM model, the ‡ indicates that in the Perception Module, the summary statistics were not transformed into to space of empirical emotion judgments.

Appendix A

Appendix to Chapter 3

A.1 Limitations

Basing our experimental stimuli on a real-world, televised gameshow increases the ecological validity of our results, but also presents methodological challenges. Our experimental stimuli were constrained by what actually happened on the gameshow, rather than being systematically varied. In particular, the pot-sizes were not balanced between outcomes, presenting an experimental confound. However, our results show that the variance in emotion ratings is much larger between outcomes than between pot-sizes, particularly excluding the two smallest pot-sizes (3.5 and 30 USD), which likely so small that they were conceptually dissimilar to the larger pot sizes.

Although our videos captured spontaneous, rich expressions, their televised nature still differs from everyday emotional situations. First, our videos did not capture the entirety of the expressions produced by the players. The videos do not depict fully temporally contiguous expressions, since they were limited by the clips that were chosen to be included on the show. The episodes of GoldenBalls typically switch between views of a player and other material, so we edited together several sections surrounding the reveal in order to gather video of five seconds featuring a player's expressions.

The contextual information was also not naturalistic in some respects. We required

that the space of events could be well-characterized in order for us to elucidate the effect of context. However, it is unclear how these results apply to everyday situations where contexts can be more complex rather than discrete possibilities. It should be noted that players on the game likely experienced a wide variety of considerations that are not made explicit by the four outcome categories, such as the monetary impacts of winning or losing given their personal financial situation, or their personal social/reputational impacts of defecting in a public game.

One limitation of our data analysis and Bayesian model was our focus on average emotion predictions across the population and across experiment stimuli, rather than on a trial-by-trial basis. Anzellotti et al. (2021) showed that individual observers represent a distribution over plausible emotions, and transform that distribution into a judgment using a complex decision rule. Here, we use observers' explicit judgments to assemble a population-level distribution, but we acknowledge that the population-level distribution need not be equivalent to any individual's latent representations. Indeed, our informal data exploration revealed reliable individual differences, and we believe that not everyone shares the same intuitive theory. Exploring such differences is likely to be a promising direction for future work.

A.2 Supplementary Analyses

A.2.1 Statistics of actual gameplay

In the 287 episodes provided by Endemol UK, players cooperated 53.0% of the time (N=574). We found no evidence that players' decisions contained information about their opponents' decisions. From the observed frequency of cooperation, we compared the observed outcome frequencies to that predicted of independent decisions. Given the observed frequency of cooperating ($P = 53.0\%$) and defecting ($Q = 1 - P$), the Hardy-Weinberg equilibrium gives the ratios of outcomes that would be expected if the decisions of the two players in a dyad were independent: $p^2 + 2pq + q^2 = 1$. A chi-square test indicated the distribution expected assuming independent assortment

is indistinguishable from the observed distribution, which suggests that the decision of one player in a dyad was statistically independent of the other player's decision, $\chi_1^2 = 0.014$, $p = 0.905$.

A.2.2 Similarity structure of emotion judgments

The similarity matrices in Figure A-1 show the mean Pearson's correlation between observers' emotion judgments for Studies 1, 2, and 3. We calculated the Pearson's correlation of every 20-dimensional emotion judgment vector with every other emotion judgment vector within the same outcome (top row) and with the same stimulus (bottom row). Correlations were z-transformed before averaging.

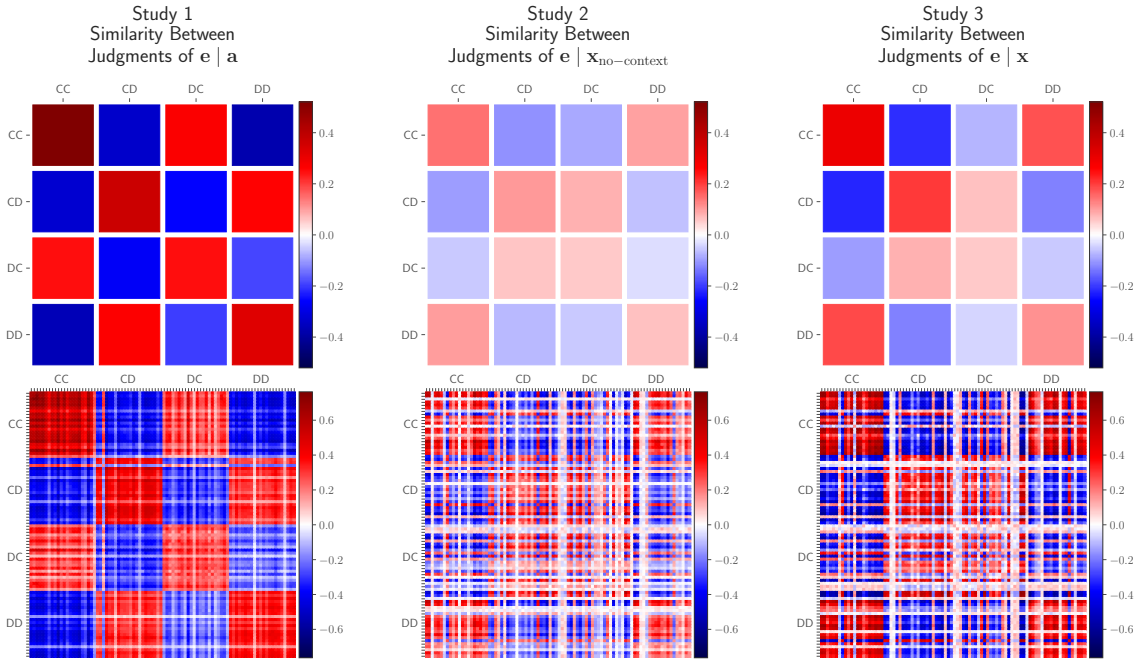

Figure A-1: Similarity between emotion judgments of the same true outcome (top row) and stimulus (bottom row).

A.2.3 PCA of emotion judgments

We performed principal component analysis on the emotion judgments from Studies 1, 2, and 3. To determine the number of significant components, we cross-validated the decomposition by censoring a random 10% of values, then determined when additional PCs reduced the fit to the held-out data. The number of PCs was 4 for each dataset.

The proportion of variance explained by the first 4 PCs, and how the PCs load onto the 20 emotions, are shown in Figure A-2. In the top row, emotions are ordered according to the amount of variance explained by the first principal component. While prior work has found that emotion judgments are effectively captured by using a two-dimensional space of 'valence' and 'arousal' (e.g. Kuppens et al., 2013; Russell, 1980), we find that four orthogonal bases are necessary to capture observers' judgments of the 20 fine-grained emotions collected. The third and fourth PCs explain only a small proportion of the overall variance, but are especially important for social emotions like *guilty* and *embarrassed* in Studies 1 and 3, and *apprehensive* and *terrified* in Study 2.

The second row of Figure A-2 shows how the first two PCs load onto each emotion. In all three studies, the first PC loads positively with the negatively-valenced emotions, and negatively with *surprised* and the positively-valenced emotions. This is consistent with the previous work on emotion judgments, which reliably finds that the most important dimension organizing emotion judgments is emotional valence (Kuppens et al., 2013; Ong et al., 2015; Russell, 1980). Unlike this previous work, the second dimension in these data does not appear to reflect emotional arousal.

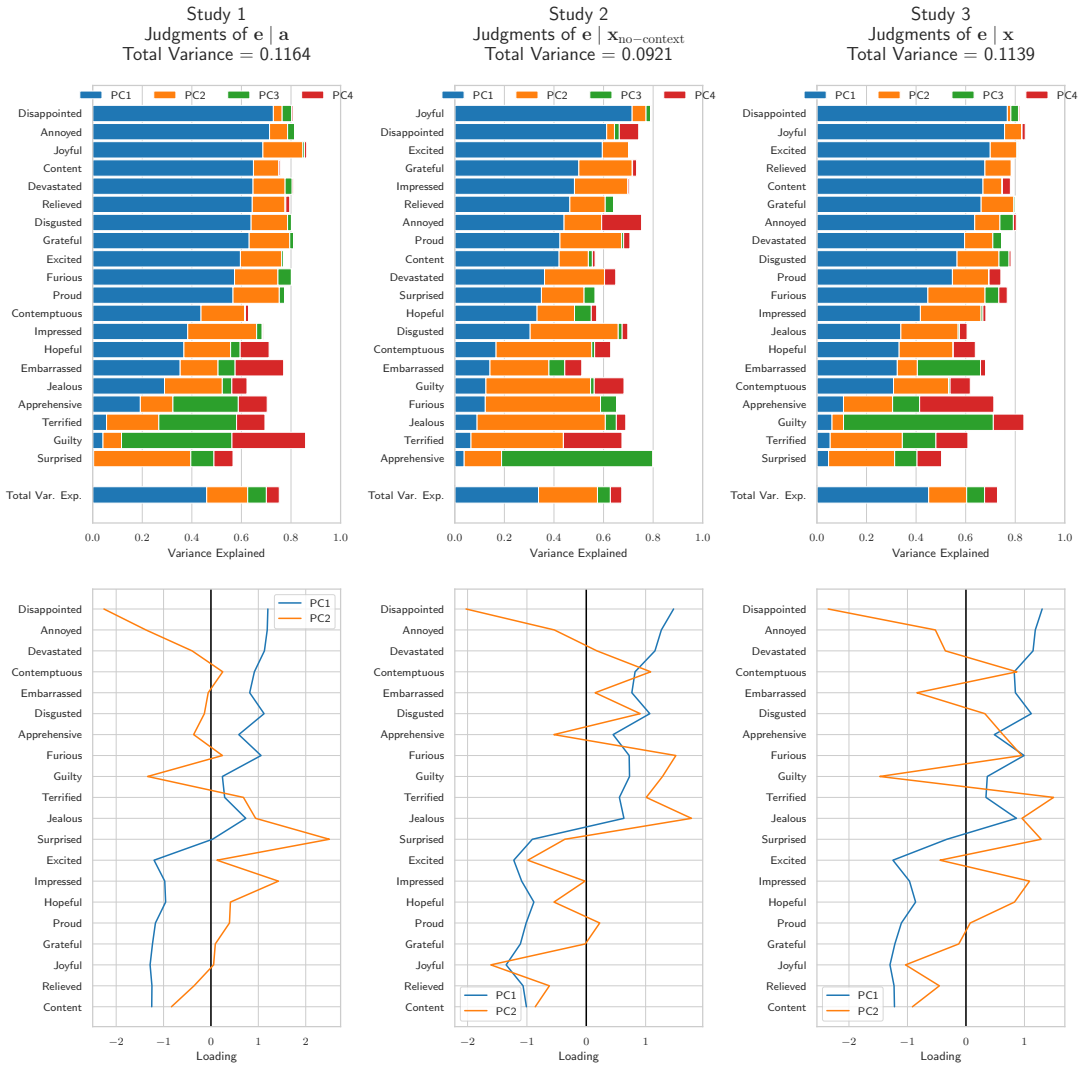

Figure A-2: Principal component analysis of emotion judgments. Top row shows the proportion of variance explained by the four PCs. Bottom row shows how the first two PCs load on to each emotion.

A.2.4 Reliability of emotion judgments

Inter-rater reliability was estimated by comparing the judgments on one observer to the mean judgments of all other observers using Pearson's correlation. Figure 3-1 gives the inter-rater reliability across emotions and players. We conducted the same analysis within emotion, within stimulus, and within outcome. Figure A-3 shows the inter-rater reliability within emotion (across players). Emotions are ordered according to the median observer reliability. Figure A-4 shows the inter-rater reliability within stimulus (across emotions). Stimuli are ordered by the outcome of the game and the size of the pot in that game. Figure A-5 shows the inter-rater reliability within outcome. This is a summary of A-4, i.e. each box plot reflects the median values of each stimulus. As in that figure, colors correspond with the outcomes.

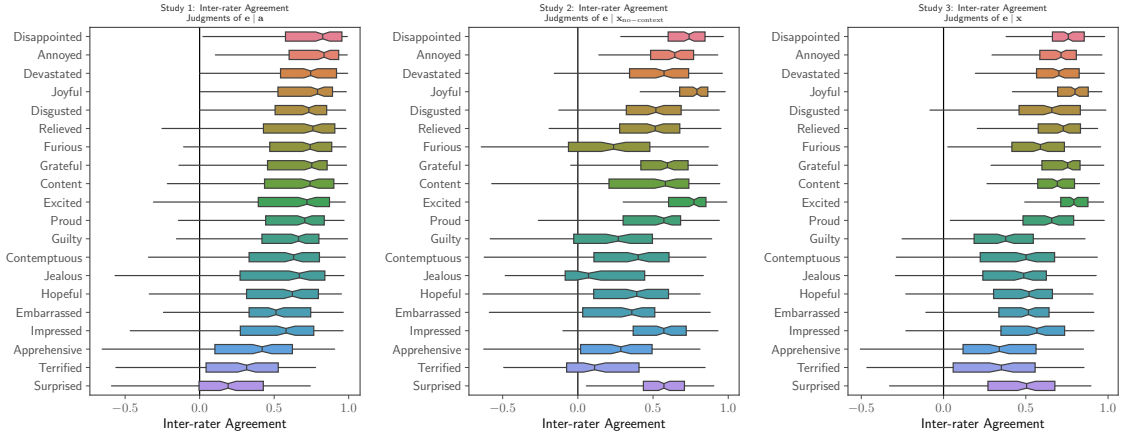

Figure A-3: Inter-rater reliability within emotion (across stimuli). Study 1 (emotions predicted from event descriptions) define the order for Study 2 and Study 3.

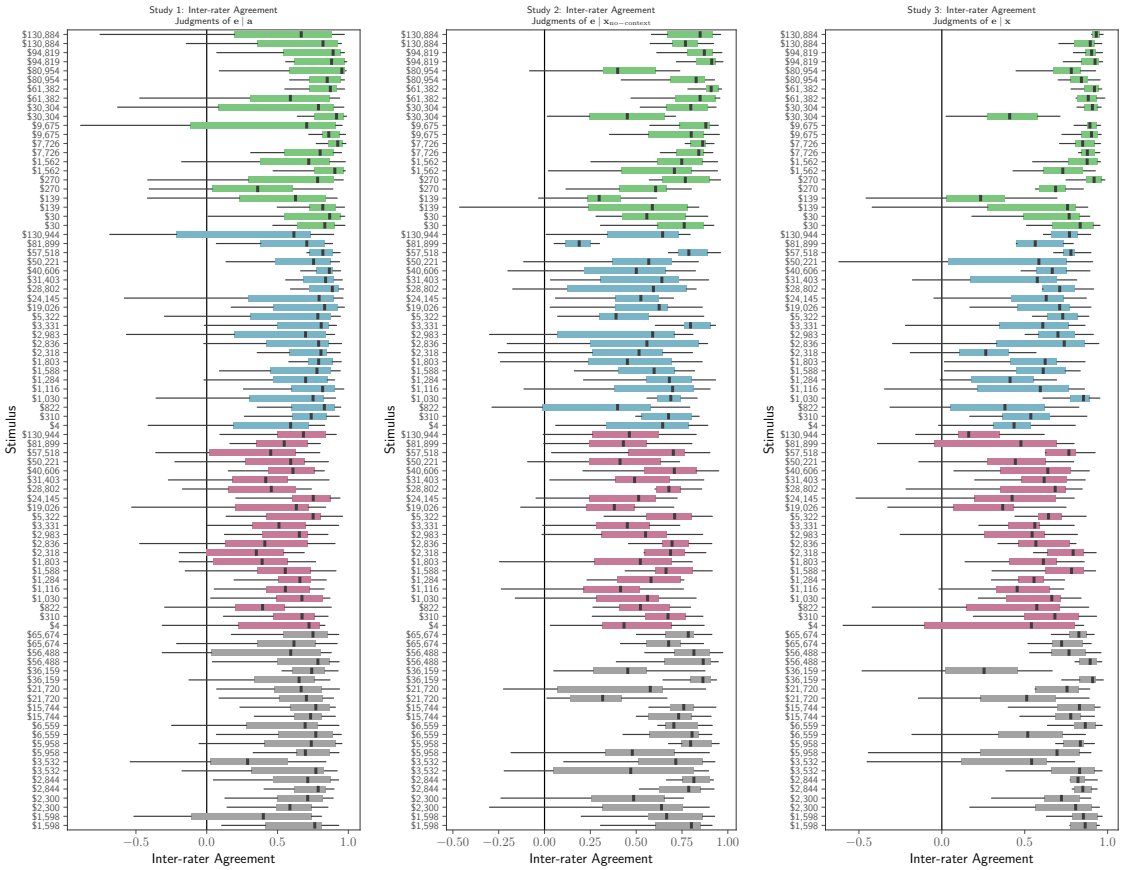

Figure A-4: Inter-rater reliability within stimulus (across emotions). Labels give the size of the pot in the games and colors indicate the outcome of the game (`CC` =green, `CD` =blue, `DC` =red, `DD` =black).

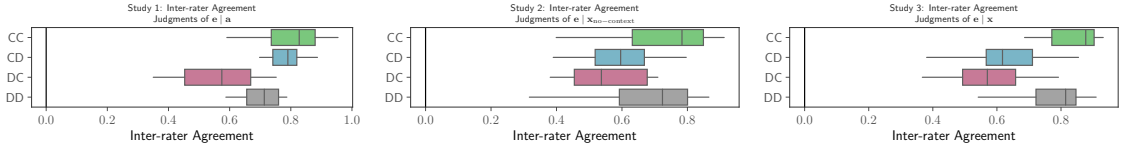

Figure A-5: Inter-rater reliability by stimulus (across emotions), aggregated by outcome.

A.2.5 Conceptual knowledge affects the interpretation of expressions

In Study 3, we found that broad-contextual knowledge about the GoldenBalls gameshow (not including what happened to the players) shifted how observers interpreted players' spontaneous dynamic expressions. Figure A-6 shows the change in the mean intensity judgments for every player's emotions. Each plot depicts 22 lines corresponding to the express videos taken from games of a given outcome. The lines connect the mean intensity attributed to the player when observers had no knowledge of the context and the mean intensity attributed to the player when observers knew that the player was on the GoldenBalls gameshow. Thus, the slope of the line reflects how broad-contextual knowledge shifted attributions of a given emotion to that player's dynamic expression. The boxplots summarize the mean intensities for all players in games of that outcome.

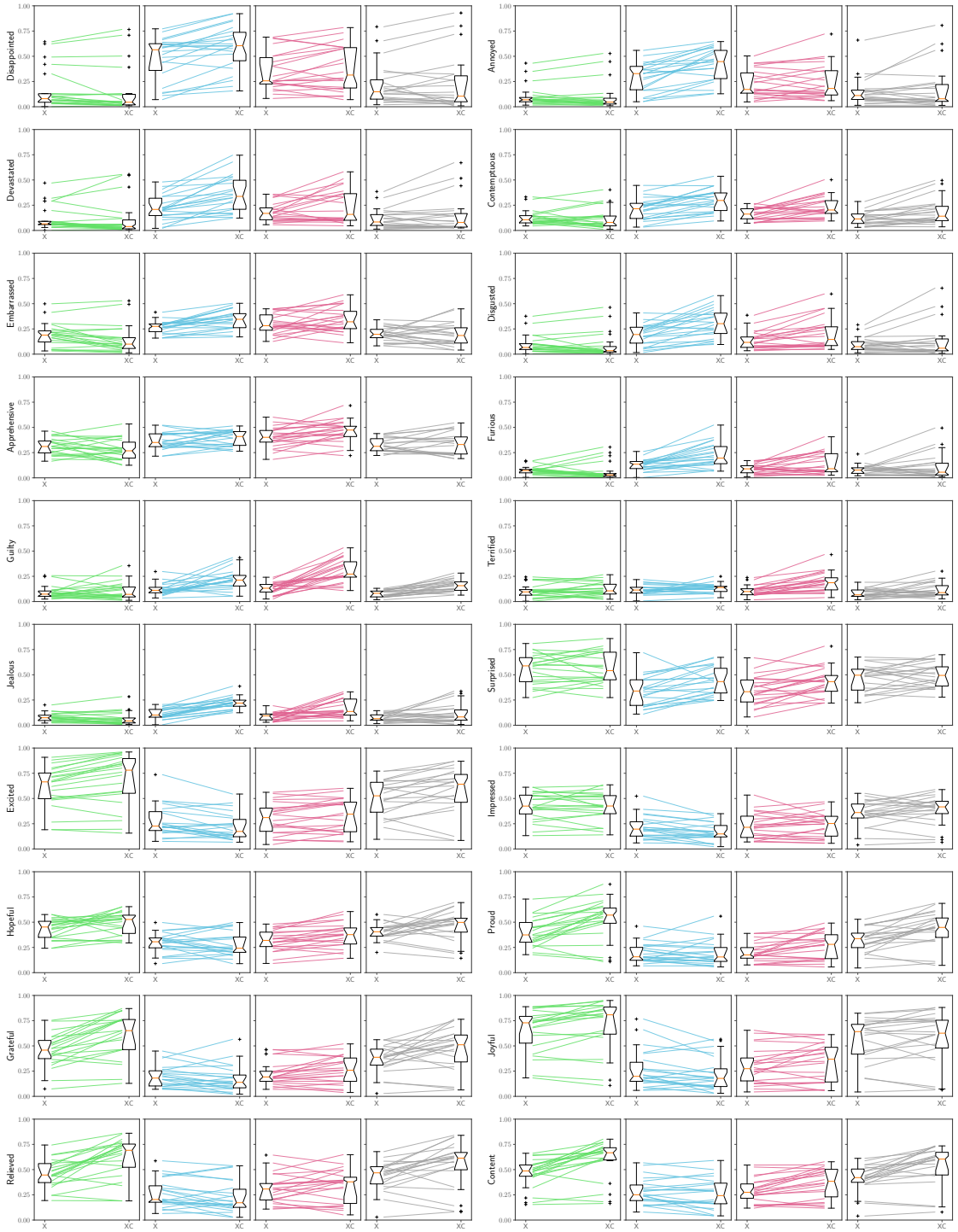

Figure A-6: Context-induced bias in the interpretation of expressions. Slope of a line illustrates how judgments of an emotion changed for that expression video when observers had informative event priors.

A.2.6 Human causal reasoning

Table A.1 summarizes observers' success at inferring the true outcome of games based on players' spontaneous dynamic expressions. We use the macro-averaged F-score as the overall score across outcomes. The F-scores of observers were tested against the chance-level F-scores that would be expected if each observer's judgments were generated based on the individual's simple response bias, independent of the stimuli, using a Wilcoxon signed-rank test. We also report the Wilcoxon statistic relative to uniform chance (0.25), which ignores the observed subject-level response biases. Figure A-7 shows the F-scores of individual observers for expression videos from each game outcome.

Table A.1: Human Causal Reasoning - Ground Truth Metrics

	Accuracy	F-score	Wilcoxon	Wilcoxon (relative to uniform chance)
Overall	0.37 [0.35,0.38]	0.35 [0.33,0.36]	$z = 7.945$, $p < 0.001$	$z = 7.562$, $p < 0.001$
CC	0.54 [0.51,0.58]	0.52 [0.49,0.54]	$z = 8.224$, $p < 0.001$	$z = 8.347$, $p < 0.001$
CD	0.51 [0.47,0.55]	0.47 [0.44,0.49]	$z = 8.201$, $p < 0.001$	$z = 7.548$, $p < 0.001$
DC	0.25 [0.22,0.28]	0.27 [0.23,0.31]	$z = 4.386$, $p < 0.001$	$z = 1.301$, $p = 0.193$
DD	0.16 [0.13,0.19]	0.15 [0.12,0.18]	$z = -4.041$, $p < 0.001$	$z = -5.910$, $p < 0.001$

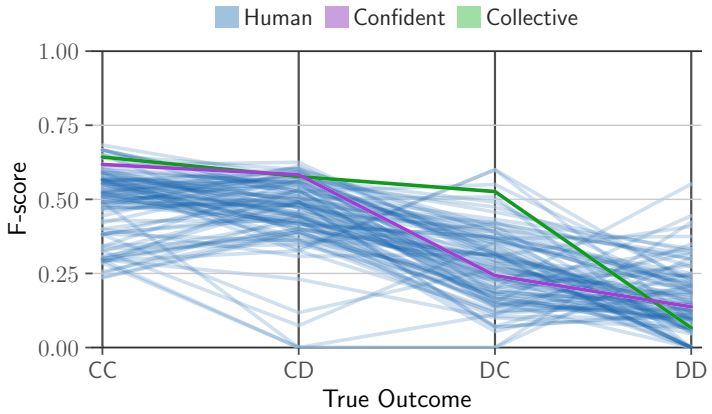

Figure A-7: Blue lines connect the four F-scores of a single observer. Purple gives F-scores of the maximally confident judgments, irrespective of the expression video. Green gives F-scores of the most popular judgment for each expression video.

Reliability of outcome judgments

In a fashion similar to Studies 1-3, we estimated the reliability of outcome judgments by correlating each observer's judgments with the population mean. To do this, the categorical outcome judgments were expressed as one-hot encoded vectors. Across all expression videos, observers' outcome judgments showed a median correlation of 0.50 [0.48, 0.51]. Fleiss' kappa = 0.203; Krippendorff's alpha = 0.203

A.2.7 Simulation of collective outcome judgments

In many domains, aggregating non-expert judgments into a population average can improve accuracy. When individual judgments are independent, noisy, and unbiased estimates of a true value, pooling judgments increases the accuracy of judgments. In this case, the collective classification accuracy increases with the number of judgments, asymptotically reaching perfect accuracy (King & Cowlishaw, 2007). If the judgments are not simply noisy, but biased, increasing the size of the pool can offer diminishing improvements to collective performance. We tested if participants showed better collective performance by simulating different group sizes, iteratively sampling participants without replacement. In each simulation, the collective judgment was determined by simple majority. Across outcomes, the collective performance shows relatively little variation with the number of participants. Larger group size improved the collective classification for primarily the `DC` outcome. Classification of the `CC` and `CD` outcomes showed only marginal improvements for group sizes larger than 15. Pooling judgments did not improve the classification of `DD` outcomes at all. For the full group, the collective macro-averaged F-score was 0.453, and the macro-averaged ROC-AUC was 0.67.

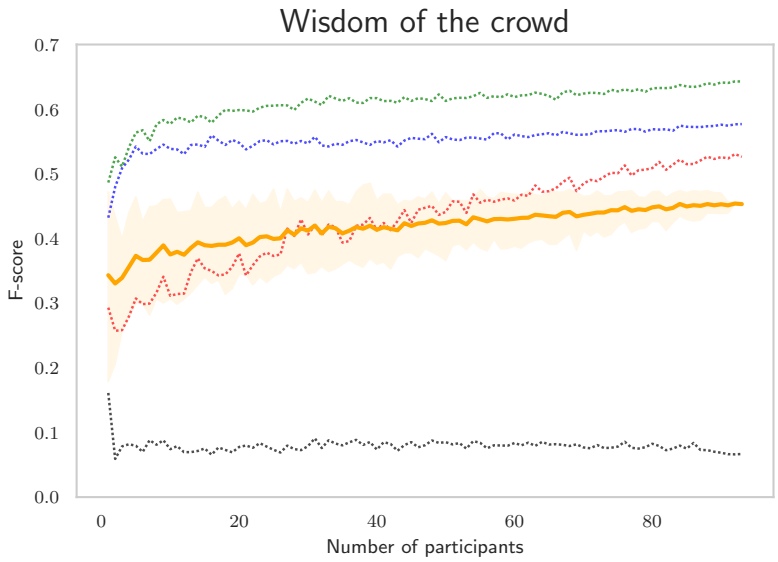

Figure A-8: Simulated group size versus collective outcome classification performance. Colors of dotted lines indicate the true outcome of the videos that were judged (CC=green; CD=blue; DC=red; DD=black). The solid yellow line gives the overall F-score (macroaveraged over outcomes) and 95% bootstrap CI.

A.2.8 Abductive inference model

In Study 5, the Bayesian belief updating model predicts the distribution of outcome judgments for each expression video, i.e. the fraction of observers that will infer that a given player's dynamic expression was a reaction to a CC game, what fraction will infer it was a reaction to a CD game, etc.

A useful way of visualizing the performance of the model results is to compare the model's predictions against human observers' judgments for every combination of inferred outcome and true outcome. The data shown in Figure 3-6c are the aggregate of the data shown in Figure A-9. Here, each plot shows the proportion of model predictions vs human judgments of the 22 videos of a given outcome. The four plots in a row show the data corresponding to the same 22 videos from games of a given outcome, with the color indicating the true outcome.

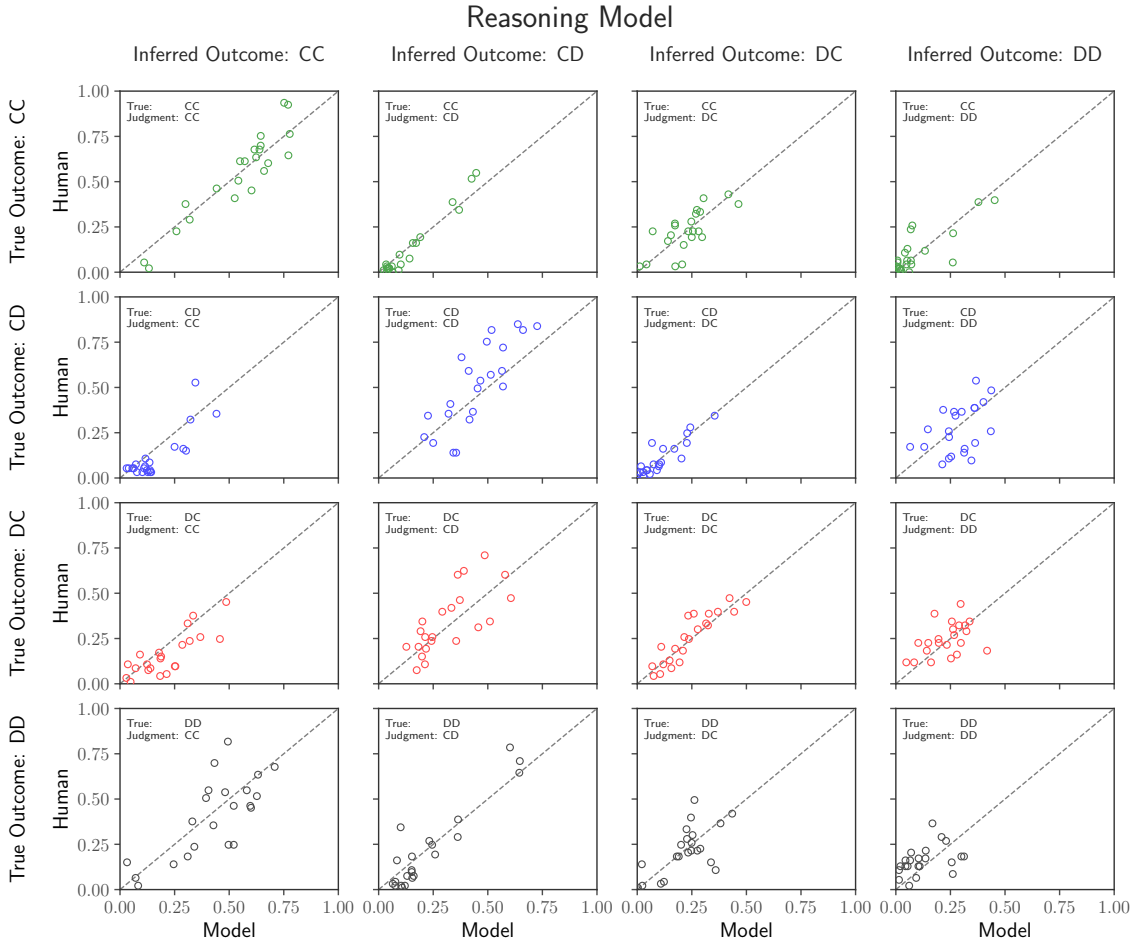

Figure A-9: *Emotion Reasoning* model predictions vs human judgments. Colors correspond to the true outcome of the stimuli. Each subplot shows 22 expression videos of a given true outcome (subplot rows), and the proportion that those videos were inferred to be a given outcome (subplot columns), by the *Emotion Reasoning* model (x-axis) and human observers (y-axis).

Appendix B

Appendix to Chapter 4

B.1 Methods

B.1.1 Mental content attribution prompts

Observers judged a player's preferences and belief on continuous scales from *"not at all"* to *"a great deal"*. For the generic players, observers rated a player's base preferences and belief, for the specific players, observers rated all seven items. Pronouns updated dynamics based on the player.

"How much does she **actually** care about..."

> ***base* Money**: "getting money (acquiring as much money as possible)"
>
> ***base* AIA**: "not getting too much (sharing with the other player, not getting more than them)"
>
> ***base* DIA**: "not getting too little (having at least as much as the other player, not getting less than them)"

"How much does he want a **reputation** for..."

> ***repu* Money**: "not prioritizing money (people believing that she values other things above maximizing her own personal financial gain)"
>
> ***repu* AIA**: "being considerate (people believing that she does not want to take advantage of her opponent)"

***repu* DIA**: "being competitive (people believing that she does not want to be taken advantage of by her opponent)"

Belief about a_2: "what does she expect the other player to choose (and how confident is she in her prediction about the other player's decision)?"

B.1.2 Utilities

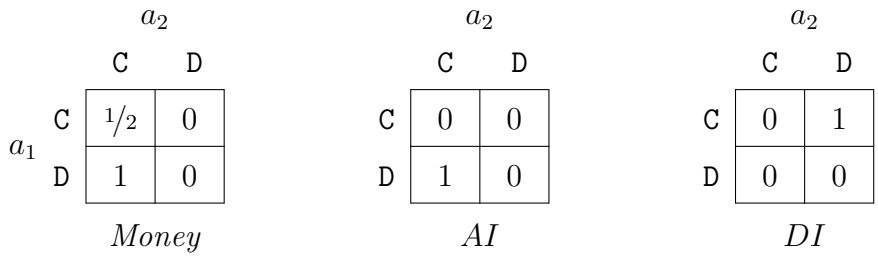

Figure B-1: **Payoff functions.** Decisions made by the two players jointly determine the players' relative payoffs in 'Split or Steal'. Payoff functions reflect how relevant the outcome is to a set of base values. The outcome's relevance is scaled by the size of the jackpot, which projects to outcome onto the dimensions of value, which are then weighted by player 1's preferences to yield subjective utilities; equation (4.1). With respect to monetary value, the payoff function simply returns the proportion of the jackpot that player 1 wins. When player 1 defects and the opponent cooperates, player 1 takes the whole pot: $Money(a_1{=}\texttt{D}, a_2{=}\texttt{C}) = 1{\cdot}pot$. Advantageous Inequity (AI) returns how much more player 1 received than player 2, and Disadvantageous Inequity (DI) returns how much more player 2 received than player 1. For the same decisions, $AI(\texttt{DC}) = 1{\cdot}pot$ and $DI(\texttt{DC}) = 0$.

The base subjective utilities are defined as functions of the monetary rewards that the two players in a dyad received. The rewards paid out to the players are, in turn, a function of the players' choices and the pot size. A simulated agent's belief about some variable x is denoted as $\pi(x)$, or equivalently, π_x.

$$\begin{aligned} \text{Let } \ \nu(x) = \text{sgn}(x) \cdot \log(1 + |x|) \ , \ \text{ and } \ \nu^{-1}(x) = \text{sgn}(x) \cdot \Big(\mathrm{e}^{|x|} - 1\Big) \\ U^{base}_{Money} = \ \ \omega^{base}_{Money} \cdot \nu\Big(Money(a_1, a_2) - \pi(Money)\Big) \\ U^{base}_{AIA} = -\omega^{base}_{AIA} \cdot \nu\Big(AI(a_1, a_2)\Big) \\ U^{base}_{DIA} = -\omega^{base}_{DIA} \cdot \nu\Big(DI(a_1, a_2)\Big) \end{aligned} \tag{B.1}$$

For each payoff function $i \in \{\, Money,\ AIA,\ DIA \,\}$, the reputation utility is defined

as:

$$U_i^{repu} = \omega_i^{repu} \cdot \nu\Big(pot \cdot \mathrm{E}_{AnonymousGame}\left[\omega_i^{base} \mid a_1\right]\Big) \tag{B.2}$$

where the sign of ω_i^{repu} is opposite that of ω_i^{base}. Note that, unlike the base utilities, the reputation utilities are not functions of a_2.

B.1.3 Planning

Anonymous Game

$$\mathrm{P}(a_1|\boldsymbol{\omega}^{base}, \pi_{a_2}) \propto \exp\Big(\lambda \sum_i \mathop{\mathrm{E}}_{a_2 \sim \pi(a_2)} U_i^{base}\Big), \qquad \text{with} \quad \lambda = 2 \tag{B.3}$$

Public Game

$$\mathrm{P}(a_1|\boldsymbol{\omega}^{base}, \boldsymbol{\omega}^{repu}, \pi_{a_2}) \propto \exp\Big(\lambda \sum_i \mathop{\mathrm{E}}_{a_2 \sim \pi(a_2)} U_i^{base} + U_i^{repu}\Big), \qquad \text{with} \quad \lambda = 2 \tag{B.4}$$

B.1.4 Appraisals

For each subject utility U, let $V = \nu^{-1}(U)$ and $T(u) = sgn(u) \cdot |u|^{0.5}$. We define the 19 appraisal features, $\boldsymbol{\psi}$, to be:

Appraisal type	**Definition**	$\tilde{U}_i^{base}$	$\tilde{U}_i^{repu}$	
Utility	$\tilde{U}_i(a_1, a_2) = T\big(V_i(a_1, a_2)\big)$	✓	✓	*(6 features)*
Prediction error	$PE_i(a_1, a_2) = T\big(V_i(a_1, a_2) - \mathrm{E}_{\pi(a_2)} V_i(a_1, a_2)\big)$	✓	✗	*(3 features)*
Counterfactual on Player 1	$CFa1_i(a_1, a_2) = T\big(V_i(\neg a_1, a_2)^{\mathrm{P}(\neg a_1)}\big)$	✓	✓	*(6 features)*
Counterfactual on Player 2	$CFa2_i(a_1, a_2) = T\big(V_i(a_1, \neg a_2)^{\pi(\neg a_2)}\big)$	✓	✗	*(3 features)*
Action prediction error	$PE_{a_2}(a_1, a_2) = \mathbb{I}[a_2{=}\mathtt{C}] - \pi(a_2{=}\mathtt{C})$			*(1 feature)*

so

$$\boldsymbol{\psi} = \langle \tilde{U}^{base,repu}_{Money,AIA,DIA}, PE^{base,repu}_{Money,AIA,DIA}, CFa1^{base,repu}_{Money,AIA,DIA}, CFa2^{base}_{Money,AIA,DIA}, PE_{a2} \rangle$$

B.2 Personalized priors

We fit priors to each player (20 *SpecificPlayers*, conditioned on whether the player chose to cooperate or defect). This is different from the BasePrior and the GenericPrior, where we integrated out the player's decision and marginalized over player identity. When modeling nonspecific players, we assumed that there was not a strong population-level bias in what actions players chose (and actual players make near-chance decisions; van den Assem et al., 2012; see Appendix A.2.1). However, the personalized descriptions of the *SpecificPlayers* may induce stronger action priors. Since observers may be more confident about which decision a stockbroker will make, for example, we use priors that are conditional on both a player's identity and decision, $\mathrm{P}\left(\boldsymbol{\omega}, \pi_{a_2} \mid a_1, player\right)$, where the priors for each planning variable is still conditionally independent. While preference and belief attributions to different *SpecificPlayers* differ considerably, the priors induce the correct decision bias in the model. Every set of preferences and beliefs attributed to players who were shown to have defected biased simulated players towards defection, with a median probability and 95% CI of 81% [73, 85]. Similarly, observers' preference and belief attributions to players who cooperated induced a cooperation bias in simulated players, 67% [62, 73]. Observers' inference of how these latent mental variables generate decisions in this game are systematic, sensitive to the picture and description, and congruent with the generative process of the Inferred Appraisals model.

Bibliography

Abdíc, I., Fridman, L., McDuff, D., Marchi, E., Reimer, B., & Schuller, B. (2016). Driver frustration detection from audio and video in the wild. *IJCAI International Joint Conference on Artificial Intelligence*, 1354–1360.

Adolphs, R. (2016). How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience*, nsw153–8. https://doi.org/10.1093/scan/nsw153

Albanie, S., & Vedaldi, A. (2016, September). Learning grimaces by watching TV. In E. R. H. Richard C. Wilson & W. A. P. Smith (Eds.), *Proceedings of the british machine vision conference (BMVC)* (pp. 122.1–122.12). BMVA Press. https://doi.org/10.5244/C.30.122

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. *Psychological Science*, *16*(5), 403–410. https://doi.org/10.1111/j.0956-7976.2005.01548.x

Anderson, L. C., Rice, K., Chrabaszcz, J., & Redcay, E. (2015). Tracking the Neurodevelopmental Correlates of Mental State Inference in Early Childhood. *Developmental neuropsychology*, *40*(7-8), 379–394. https://doi.org/10.1080/87565641.2015.1119836

Anzellotti, S., Houlihan, S. D., Liburd Jr., S., & Saxe, R. (2021). Leveraging facial expressions and contextual information to investigate opaque representations of emotions. *Emotion*, *21*(1), 96–107. https://doi.org/10.1037/emo0000685

Armstrong, K., Semien, R., Miller, T. C., & Glass, I. (2016). Anatomy of Doubt. WBEZ Chicago.

Atias, D., & Aviezer, H. (2021). Empathic Accuracy: Lessons from the Perception of Contextualized Real-Life Emotional Expressions. In M. Gilead & K. N. Ochsner (Eds.), *The Neural Basis of Mentalizing* (pp. 171–188). Springer International Publishing. https://doi.org/10.1007/978-3-030-51890-5_9

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, Disgusted, or Afraid?: Studies on the Malleability of Emotion Perception. *Psychological Science*, *19*(7), 724–732. https://doi.org/10.1111/j.1467-9280.2008.02148.x

Aviezer, H., Messinger, D. S., Zangvil, S., Mattson, W. I., Gangi, D. N., & Todorov, A. (2015). Thrill of victory or agony of defeat? Perceivers fail to utilize information in facial movements. *Emotion*, *15*(6), 791–797. https://doi.org/10.1037/emo0000073

Aviezer, H., Trope, Y., & Todorov, A. (2012a). Holistic person processing: Faces with bodies tell the whole story. *Journal of Personality and Social Psychology*, *103*(1), 20–37. https://doi.org/10.1037/a0027411

Aviezer, H., Trope, Y., & Todorov, A. (2012b). Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science*, *338*(6111), 1225–1229. https://doi.org/10.1126/science.1224313

B. Martinez, M. F. Valstar, B. Jiang, & M. Pantic. (2019). Automatic Analysis of Facial Actions: A Survey. *IEEE Transactions on Affective Computing*, *10*(3), 325–347. https://doi.org/10.1109/TAFFC.2017.2731763

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 598. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Balaz, V., Bačová, V., Drobná, E., Dudeková, K., & Adamík, K. (2013). Testing prospect theory parameters. *Ekonomicky casopis*, *61*(7), 655–671.

Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. *Emotion Review*, *6*(4), 292–297. https://doi.org/10.1177/1754073914534479

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

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 68. https://doi.org/10.1177/1529100619832930

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in Emotion Perception. *Current directions in psychological science*, *20*(5), 286–290. https://doi.org/10.1177/0963721411422522

Bartlett, M. S., Littlewort, G. C., Frank, M. G., & Lee, K. (2014). Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions. *Current Biology*, *24*(7), 738–743. https://doi.org/10.1016/j.cub.2014.02.009

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*(2), 170–176.

Battigalli, P., Dufwenberg, M., & Smith, A. (2015, March 30). *Frustration and Anger in Games* (SSRN Scholarly Paper No. ID 2591839). Social Science Research Network. Rochester, NY.

Baudouin, J. Y., Sansone, S., & Tiberghien, G. (2000). Recognizing expression from familiar and unfamiliar faces. *Pragmatics & Cognition*, *8*(1), 123–146. https://doi.org/10.1075/pc.8.1.07bau

Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion*, *19*(8), 1463–1477. https://doi.org/10.1037/emo0000510

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W.-t., & Choi, Y. (2019). Abductive commonsense reasoning. https://doi.org/10.48550/ARXIV.1908.05739

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., Wrangham, R., & Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*(7581), 258–261. https://doi.org/10.1038/nature15703

Böhm, G., & Pfister, H.-R. (2015). How people explain their own and others' behavior: A theory of lay causal explanations. *Frontiers in psychology*, *6*, 139. https://doi.org/10.3389/fpsyg.2015.00139

Bolton, G. E., & Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, *90*(1), 166–193. https://doi.org/10.1257/aer.90.1.166

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2013). The modular nature of trustworthiness detection. *Journal of experimental psychology. General*, *142*(1), 143–150. https://doi.org/10.1037/a0028930

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. https://doi.org/10.18653/v1/P19-1470

Brooks, J. A., Chikazoe, J., Sadato, N., & Freeman, J. B. (2019). The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences*, *116*(32), 15861–15870. https://doi.org/10.1073/pnas.1816408116

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature Human Behaviour*, 1–14. https://doi.org/10.1038/s41562-018-0376-6

Brooks, J. A., Stolier, R. M., & Freeman, J. B. (2018). Stereotypes Bias Visual Prototypes for Sex and Emotion Categories. *Social Cognition*, *36*(5), 481–493. https://doi.org/10.1521/soco.2018.36.5.481

Bryant, D., & Howard, A. (2019). A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity.

Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 377–382. https://doi.org/10.1145/3306618.3314284

Buck, R. (1994). Social and emotional functions in facial expression and communication: The readout hypothesis. *Biological Psychology*, *38*(2), 95–115. https://doi.org/10.1016/0301-0511(94)90032-9

Buck, R., Losow, J. I., Murphy, M. M., & Costanzo, P. (1992). Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology*, *63*(6), 962–968. https://doi.org/10.1037/0022-3514.63.6.962

Burton-Chellew, M. N., & West, S. A. (2012). Correlates of Cooperation in a One-Shot High-Stakes Televised Prisoners' Dilemma. *PLOS ONE*, *7*(4), e33344. https://doi.org/10.1371/journal.pone.0033344

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Carey, S. (2009, May). *The Origin of Concepts*. Oxford University Press.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, *70*(2), 205–218. https://doi.org/10.1037/0022-3514.70.2.205

Chen, J., Yang, T., Huang, Z., Wang, K., Liu, M., & Lyu, C. (2022). Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03729-4

Chen, Y., Yang, X., Cham, T.-J., & Cai, J. (2021). Towards Unbiased Visual Emotion Recognition via Causal Intervention. https://doi.org/10.48550/ARXIV.2107.12096

Chen, Z., & Whitney, D. (2019). Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, *116*(15), 7559–7564. https://doi.org/10.1073/pnas.1812250116

Chen, Z., & Whitney, D. (2020). Inferential emotion tracking (IET) reveals the critical role of context in emotion recognition. *Emotion*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/emo0000934

Chovil, N. (1991). Social determinants of facial displays. *Journal of Nonverbal Behavior*, *15*(3), 141–154. https://doi.org/10.1007/BF01672216

Coleman, M. D. (2016). Emotion and the False Consensus Effect. *Current Psychology*, 1–7. https://doi.org/10.1007/s12144-016-9489-0

Cooney, G., Gilbert, D. T., & Wilson, T. D. (2014). The unforeseen costs of extraordinary experience. *Psychological Science*, *25*(12), 2259–2265. https://doi.org/10.1177/0956797614551372

Cordaro, D. T., Sun, R., Kamble, S., Hodder, N., Monroy, M., Cowen, A., Bai, Y., & Keltner, D. (2020). The recognition of 18 facial-bodily expressions across nine cultures. *Emotion*, *20*(7), 1292–1300. https://doi.org/10.1037/emo0000576

Cowen, A. S., & Keltner, D. (2020). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, *75*(3), 349–364. https://doi.org/10.1037/amp0000488

Crivelli, C., Carrera, P., & Fernández-Dols, J.-M. (2015). Are smiles a sign of happiness? Spontaneous expressions of judo winners. *Evolution and Human Behavior*, *36*(1), 52–58. https://doi.org/10.1016/j.evolhumbehav.2014.08.009

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, *60*(23), 685–700.

De Bruyn, A., & Bolton, G. E. (2008). Estimating the Influence of Fairness on Bargaining Behavior. *Management Science*, *54*(10), 1774–1791. https://doi.org/10.1287/mnsc.1080.0887

de Gelder, B., de Borst, A. W., & Watson, R. (2015). The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 149–158. https://doi.org/10.1002/wcs.1335

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(1), 73–88. https://doi.org/10.1037/a0034251

de Melo, C. M., Gratch, J., Carnevale, P. J., & Read, S. J. (2012). Reverse appraisal: The importance of appraisals for the effect of emotion displays on peoples decision making in a social dilemma. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, *34*, 270–275.

Do, N.-T., Kim, S.-H., Yang, H.-J., Lee, G.-S., & Yeom, S. (2021). Context-Aware Emotion Recognition in the Wild Using Spatio-Temporal and Temporal-Pyramid Models. *Sensors*, *21*(7). https://doi.org/10.3390/s21072344

Dobs, K., Bülthoff, I., Breidt, M., Vuong, Q. C., Curio, C., & Schultz, J. (2014). Quantifying human sensitivity to spatio-temporal information in dynamic faces. *Vision research*, *100*, 78–87. https://doi.org/10.1016/j.visres.2014.04.009

Doyle, C. M., Gendron, M., & Lindquist, K. A. (2021). Language Is a Unique Context for Emotion Perception. *Affective Science*, *2*(2), 171–177. https://doi.org/10.1007/s42761-020-00025-7

Dufwenberg, M., & Gneezy, U. (2000). Measuring Beliefs in an Experimental Lost Wallet Game. *Games and Economic Behavior*, *30*(2), 163–182. https://doi.org/10.1006/game.1999.0715

Dupré, D., Krumhuber, E. G., Küster, D., & McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLOS ONE*, *15*(4), e0231968. https://doi.org/10.1371/journal.pone.0231968

Durán, J. I., & Fernández-Dols, J.-M. (2021). Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. *Emotion*, *21*(7), 1550–1569. https://doi.org/10.1037/emo0001015

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guide-lines for research and an integration of findings*. Pergamon Press.

Elfenbein, H. A., & Ambady, N. (2003). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology*, *85*(2), 276–290. https://doi.org/10.1037/0022-3514.85.2.276

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In *Handbook of affective sciences* (pp. 572–595). Oxford University Press.

Engelmann, J. M., & Tomasello, M. (2019). Children's Sense of Fairness as Equal Respect. *Trends in Cognitive Sciences*, *23*(6), 454–463. https://doi.org/10.1016/j.tics.2019.03.001

Evans, O., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 323–329.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry*, *41*(1), 20–26. https://doi.org/10.1093/ei/41.1.20

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Fernberger, S. W. (1928). False Suggestion and the Piderit Model. *The American Journal of Psychology*, *40*(4), 562. https://doi.org/10.2307/1414334

Ferrari, C., Lega, C., Vernice, M., Tamietto, M., Mende-Siedlecki, P., Vecchi, T., Todorov, A., & Cattaneo, Z. (2016). The Dorsomedial Prefrontal Cortex Plays a Causal Role in Integrating Social Impressions from Faces and Verbal Descriptions. *Cerebral Cortex*, *26*(1), 156–165. https://doi.org/10.1093/cercor/bhu186

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*(2), 229–240. https://doi.org/10.1037/0022-3514.60.2.229

Gao, J., Qing, L., Li, L., Cheng, Y., & Peng, Y. (2021). Multi-scale features based interpersonal relation recognition using higher-order graph neural network. *Neurocomputing*, *456*, 243–252. https://doi.org/10.1016/j.neucom.2021.05.097

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. https://doi.org/10.1016/S1364-6613(03)00128-1

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016). Plans, Habits, and Theory of Mind. *PLoS ONE*, *11*(9), e0162246–24. https://doi.org/10.1371/journal.pone.0162246

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive Theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, *75*(3), 617–638. https://doi.org/10.1037/0022-3514.75.3.617

Goel, S., Jara-Ettinger, J., & Gendron, M. (2022). Modeling Cue-integration in Emotion Inferences. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, *44*, 862–868.

Goldenberg, A., Schöne, J., Huang, Z., Sweeny, T. D., Ong, D. C., Brady, T. F., Robinson, M. M., Levari, D., Zaki, J., & Gross, J. J. (2022). Amplification in the evaluation of multiple emotional expressions over time. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01390-y

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality (2010/02/04). *Behavioral and Brain Sciences*, *16*(1), 1–14. https://doi.org/10.1017/S0140525X00028636

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Goren, A., & Todorov, A. (2009). Two faces are better than one: Eliminating false trait associations with faces. *Social Cognition*, *27*(2), 222–248. https://doi.org/10.1521/soco.2009.27.2.222

Gratch, J., & Marsella, S. C. (2014, December). Appraisal Models. In *The oxford handbook of affective computing* (pp. 54–67). Oxford University Press, USA.

Haidt, J., & Keltner, D. (1999). Culture and Facial Expression: Open-ended Methods Find More Expressions and a Gradient of Recognition. *Cognition and Emotion*, *13*(3), 225–266. https://doi.org/10.1080/026999399379267

Harrison, P. M. C., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *Advances in Neural Information Processing Systems*, *33*. https://doi.org/10.17605/OSF.IO/RZK4S

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently Ambiguous: Facial Expressions of Emotions, in Context. *Emotion Review*, *5*(1), 60–65. https://doi.org/10.1177/1754073912451331

Hayashi, N., Ostrom, E., Walker, J., & Yamagishi, T. (1999). RECIPROCITY, TRUST, AND THE SENSE OF CONTROL: A CROSS-SOCIETAL STUDY. *Rationality and Society*, *11*(1), 27–46. https://doi.org/10.1177/104346399011001002

Hehman, E., Flake, J. K., & Freeman, J. B. (2015). Static and Dynamic Facial Cues Differentially Affect the Consistency of Social Evaluations. *Personality and Social Psychology Bulletin*, *41*(8), 1123–1134. https://doi.org/10.1177/0146167215591495

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243. https://doi.org/10.2307/1416950

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815. https://doi.org/10.1017/S0140525X05000142

Hess, U., Adams, R., & Kleck, R. (2005). Who may frown and who should smile? Dominance, affiliation, and the display of happiness and anger. *Cognition and Emotion*, *19*(4), 515–536. https://doi.org/10.1080/02699930441000364

Hess, U., Banse, R., & Kappas, A. (1995). The intensity of facial expression is determined by underlying affective state and social situation. *Journal of Personality and Social Psychology*, *69*(2), 280–288. https://doi.org/10.1037/0022-3514.69.2.280

Hess, U., & Hareli, S. (2015). The influence of context on emotion recognition in humans. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, *03*, 1–6. https://doi.org/10.1109/FG.2015.7284842

Hoemann, K., Crittenden, A. N., Msafiri, S., Liu, Q., Li, C., Roberson, D., Ruark, G. A., Gendron, M., & Feldman Barrett, L. (2019). Context facilitates performance on a classic cross-cultural emotion perception task. *Emotion*, *19*(7), 1292–1313. https://doi.org/10.1037/emo0000501

Houlihan, S. D., Kleiman-Weiner, M., Tenenbaum, J. B., & Saxe, R. (2018). A generative model of people's intuitive theory of emotions: Inverse planning in rich social games. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, *40*.

Houlihan, S. D., Ong, D., Cusimano, M., & Saxe, R. (2022). Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory

of mind. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, *44*, 854–861.

Houlihan, S. D., Tenenbaum, J. B., & Saxe, R. (2021). Linking Models of Theory of Mind and Measures of Human Brain Activity. In M. Gilead & K. N. Ochsner (Eds.), *The Neural Basis of Mentalizing* (pp. 209–235). Springer International Publishing. https://doi.org/10.1007/978-3-030-51890-5_11

Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2021). (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 6384–6392.

Ichimura, T., & Kamada, S. (2022). An Ensemble Learning Method of Adaptive Structural Deep Belief Network for AffectNet. *International Journal of Smart Computing and Artificial Intelligence*, *6*(1), 1. https://doi.org/10.52731/ijscai.v6.i1.640

Israelashvili, J., Hassin, R. R., & Aviezer, H. (2019). When emotions run high: A critical role for context in the unfolding of dynamic, real-life facial affect. *Emotion*, *19*(3), 558–562. https://doi.org/10.1037/emo0000441

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*(2), 288–299. https://doi.org/10.1037/0033-2909.115.2.288

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23. https://doi.org/10.1016/j.cognition.2015.03.006

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696–9701. https://doi.org/10.1073/pnas.1719452115

Jern, A., & Kemp, C. (2015). A decision network account of reasoning about other peoples choices. *Cognition*, *142*, 12–38. https://doi.org/10.1016/j.cognition.2015.05.006

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64. https://doi.org/10.1016/j.cognition.2017.06.017

Jin, B. T., Abdelrahman, L., Chen, C. K., & Khanzada, A. (2020). Fusical: Multimodal fusion for video sentiment. *Proceedings of the 2020 International Conference on Multimodal Interaction*, 798–806.

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

Kayyal, M., Widen, S., & Russell, J. A. (2015). Context is more powerful than we think: Contextual cues override facial cues even for valence. *Emotion*, *15*(3), 287–291. https://doi.org/10.1037/emo0000032

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional Expression: Advances in Basic Emotion Theory. *Journal of Nonverbal Behavior*, *43*(2), 133–160. https://doi.org/10.1007/s10919-019-00293-3

Khaireddin, Y., & Chen, Z. (2021, May 8). Facial Emotion Recognition: State of the Art Performance on FER2013.

King, A. J., & Cowlishaw, G. (2007). When to use social information: The advantage of large group size in individual decision making. *Biology Letters*, *3*(2), 137–139. https://doi.org/10.1098/rsbl.2007.0017

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior*, *21*(6), 411–427. https://doi.org/10.1016/S1090-5138(00)00055-6

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, *39*, 676–681.

Kliemann, D., & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology*, *24*, 1–6.

Kosti, R., Alvarez, J., Recasens, A., & Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion Recognition in Context. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1960–1968. https://doi.org/10.1109/cvpr.2017.212

Kotz, S. A., Kalberlah, C., Bahlmann, J., Friederici, A. D., & Haynes, J.-D. (2012). Predicting vocal emotion expressions from the human brain. *Human Brain Mapping*, *34*(8), 1971–1981. https://doi.org/10.1002/hbm.22041

Kruger, J., & Gilovich, T. (1999). "Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, *76*(5), 743–753. https://doi.org/10.1037/0022-3514.76.5.743

Krumhuber, E. G., & Scherer, K. R. (2016). The Look of Fear from the Eyes Varies with the Dynamic Sequence of Facial Actions. *Swiss Journal of Psychology*, *75*(1), 5–14. https://doi.org/10.1024/1421-0185/a000166

Krumhuber, E. G., Kappas, A., & Manstead, A. S. R. (2013). Effects of Dynamic Aspects of Facial Expressions: A Review. *Emotion Review*, *5*(1), 41–46. https://doi.org/10.1177/1754073912451349

Krumhuber, E. G., Küster, D., Namba, S., & Skora, L. (2021). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior Research Methods*, *53*(2), 686–701. https://doi.org/10.3758/s13428-020-01443-y

Kryven, M., Ullman, T., Cowan, W., & Tenenbaum, J. B. (2016). Outcome or Strategy? A Bayesian Model of Intelligence Attribution. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, *38*.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*(4), 917–940. https://doi.org/10.1037/a0030811

Kyriakou, K., Kleanthous, S., Otterbacher, J., & Papadopoulos, G. A. (2020). Emotion-based Stereotypes in Image Analysis Services. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 252–259. https://doi.org/10.1145/3386392.3399567

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Le, T. A., Collins, K. M., Hewitt, L., Ellis, K., N, S., Gershman, S. J., & Tenenbaum, J. B. (2021, July 3). *Hybrid Memoised Wake-Sleep: Approximate Inference at the Discrete-Continuous Interface*.

Le Mau, T. (2019). *Towards understanding facial movements in real life*.

Le Mau, T., Hoemann, K., Lyons, S. H., Fugate, J. M. B., Brown, E. N., Gendron, M., & Barrett, L. F. (2021). Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature Communications*, *12*(1), 5037. https://doi.org/10.1038/s41467-021-25352-6

Lecker, M., & Aviezer, H. (2021). More than Words? Semantic Emotion Labels Boost Context Effects on Faces. *Affective Science*, *2*(2), 163–170. https://doi.org/10.1007/s42761-021-00043-z

Lecker, M., Dotsch, R., Bijlstra, G., & Aviezer, H. (2020). Bidirectional contextual influence between faces and bodies in emotion perception. *Emotion*, *20*(7), 1154–1164. https://doi.org/10.1037/emo0000619

Lee, K. H., & Siegle, G. J. (2014). Different brain activity in response to emotional faces alone and augmented by contextual information. *Psychophysiology*, *51*(11), 1147–1157. https://doi.org/10.1111/psyp.12254

Lei, S., & Gratch, J. (2019). Smiles Signal Surprise in a Social Dilemma. *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 627–633. https://doi.org/10.1109/ACII.2019.8925494

Liang, P. P., Lyu, Y., Fan, X., Mo, S., Yogatama, D., Morency, L.-P., & Salakhutdinov, R. (2022, March 3). HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. https://doi.org/10.48550/ARXIV.2203.01311

Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, *45*(1), 255–268. https://doi.org/10.2307/2532051

Lindquist, K. A., Satpute, A. B., & Gendron, M. (2015). Does Language Do More Than Communicate Emotion? *Current directions in psychological science*, *24*(2), 99–108. https://doi.org/10.1177/0963721414553440

Lindquist, K. A., & Gendron, M. (2013). What's in a Word? Language Constructs Emotion Perception. *Emotion Review*, *5*(1), 66–71. https://doi.org/10.1177/1754073912451351

Lindquist, K. A., Gendron, M., Barrett, L. F., & Dickerson, B. C. (2014). Emotion perception, but not affect perception, is impaired with semantic memory loss. *Emotion*, *14*(2), 375–387. https://doi.org/10.1037/a0035293

Lombrozo, T. (2012). Explanation and abductive inference. In *The Oxford handbook of thinking and reasoning.* (pp. 260–276). Oxford University Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, *4*(3), e4638–14. https://doi.org/10.1371/journal.pone.0004638

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as Multipurpose Social Signals. *Trends in Cognitive Sciences*, *21*(11), 864–877. https://doi.org/10.1016/j.tics.2017.08.007

Martinez, L., Falvello, V. B., Aviezer, H., & Todorov, A. (2015). Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cognition & Emotion*, *30*(5), 939–952. https://doi.org/10.1080/02699931.2015.1035229

Matsumoto, D., & Hwang, H. C. (2018). Microexpressions Differentiate Truths From Lies About Future Malicious Intent. *Frontiers in Psychology*, *9*, 2545. https://doi.org/10.3389/fpsyg.2018.02545

McDuff, D., Amr, M., & Kaliouby, R. el. (2019). AM-FED+: An Extended Dataset of Naturalistic Facial Expressions Collected in Everyday Settings. *IEEE Transactions on Affective Computing*, *10*(1), 7–17. https://doi.org/10.1109/TAFFC.2018.2801311

Mehu, M., & Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, *15*(6), 798–811. https://doi.org/10.1037/a0039416

Melo, C. D., Marsella, S., & Gratch, J. (2016). People Do Not Feel Guilty About Exploiting Machines. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(2), 8:1–8:17. https://doi.org/10.1145/2890495

Miloyan, B., & Suddendorf, T. (2015). Feelings of the future. *Trends in Cognitive Sciences*, *19*(4), 196–200. https://doi.org/10.1016/j.tics.2015.01.008

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14222–14231. https://doi.org/10.1109/CVPR42600.2020.01424

Mittal, T., Mathur, P., Bera, A., & Manocha, D. (2021). Affect2MM: Affective analysis of multimedia content using emotion causality. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5661–5671.

Moors, A. (2014). Flavors of Appraisal Theories of Emotion. *Emotion Review*, *6*(4), 303–307. https://doi.org/10.1177/1754073914534477

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Consciousness and Cognition*, *25*(100), 67–76. https://doi.org/10.1016/j.concog.2014.01.009

Mumenthaler, C., & Sander, D. (2015). Automatic integration of social information in emotion recognition. *Journal of experimental psychology. General*, *144*(2), 392–399. https://doi.org/10.1037/xge0000059

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289–316.

Nakamura, M., Buck, R., & Kenny, D. A. (1990). Relative contributions of expressive behavior and contextual information to the judgment of the emotional state of another. *Journal of Personality and Social Psychology*, *59*(5), 1032–1039. https://doi.org/10.1037/0022-3514.59.5.1032

Nelson, J. A., de Lucca Freitas, L. B., OBrien, M., Calkins, S. D., Leerkes, E. M., & Marcovitch, S. (2012). Preschool-aged children's understanding of gratitude: Relations with emotion and mental state knowledge. *British Journal of Developmental Psychology*, *31*(1), 42–56. https://doi.org/10.1111/j.2044-835X.2012.02077.x

Newen, A., Welpinghus, A., & Juckel, G. (2015). Emotion Recognition as Pattern Recognition: The Relevance of Perception: Emotion Recognition as Pattern Recognition. *Mind & Language*, *30*(2), 187–208. https://doi.org/10.1111/mila.12077

O'Brien, M., Miner Weaver, J., Nelson, J. A., Calkins, S. D., Leerkes, E. M., & Marcovitch, S. (2011). Longitudinal associations between children's understanding of emotions and theory of mind. *Cognition & Emotion*, *25*(6), 1074–1086. https://doi.org/10.1080/02699931.2010.518417

Ong, D. C. (2021, July 28). *An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence*.

Ong, D. C., Asaba, M., & Gweon, H. (2016). Young children and adults integrate past expectations and current outcomes to reason about others' emotions. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, *38*, 135–140.

Ong, D. C., Soh, H., Zaki, J., & Goodman, N. D. (2021). Applying Probabilistic Programming to Affective Computing. *IEEE Transactions on Affective Computing*, *12*(2), 306–317. https://doi.org/10.1109/TAFFC.2019.2905211

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162. https://doi.org/10.1016/j.cognition.2015.06.010

Ong, D. C., Zaki, J., & Goodman, N. D. (2016). Emotions in lay explanations of behavior. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, *38*, 360–365.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational models of emotion inference in Theory of Mind: A review and roadmap. *Topics in Cognitive Science*, *11*(2), 338–357. https://doi.org/10.1111/tops.12371

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Ornaghi, V., & Grazzani, I. (2013). The relationship between emotional-state language and emotion understanding: A study with school-age children. *Cognition & Emotion*, *27*(2), 356–366. https://doi.org/10.1080/02699931.2012.711745

Ortony, A., Clore, G. L., & Collins, A. (1990, May 25). *The Cognitive Structure of Emotions*. Cambridge University Press.

Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., & Choi, Y. (2020). VisualCOMET: Reasoning about the dynamic context of a still image. *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Parkinson, B. (2005). Do Facial Movements Express Emotions or Communicate Motives? *Personality and Social Psychology Review*, *9*(4), 278–311. https://doi.org/10.1207/s15327957pspr0904_1

Parry, G., & Vuong, Q. (2021, June 8). *Deep Affect: Using objects, scenes and facial expressions in a deep neural network to predict arousal and valence values of images* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/t9p3f

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peña, A., Serna, I., Morales, A., Fierrez, J., & Lapedriza, A. (2020). Facial expressions as a vulnerability in face recognition. https://doi.org/10.48550/ARXIV.2011.08809

Pollmann, M. M. H., & Finkenauer, C. (2009). Empathic forecasting: How do we predict other people's feelings? *Cognition and Emotion*, *23*(5), 978–1001. https://doi.org/10.1080/02699930802264895

Prabhu, V. U., & Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision? https://doi.org/10.48550/ARXIV.2006.16923

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018, July 10–15). Machine theory of mind. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 4218–4227, Vol. 80). PMLR.

Rapoport, A. (1988). Experiments with N-Person Social Traps I: Prisoner's dilemma, weak prisoner's dilemma, volunteer's dilemma, and largest number. *Journal of Conflict Resolution*, *32*(3), 457–472. https://doi.org/10.1177/0022002788032003003

Repacholi, B. M., Meltzoff, A. N., Toub, T. S., & Ruba, A. L. (2016). Infants generalizations about other peoples emotions: Foundations for trait-like attributions. *Developmental Psychology*, *52*(3), 364–378. https://doi.org/10.1037/dev0000097

Reschke, P. J., Walle, E. A., Knothe, J. M., & Lopez, L. D. (2019). The influence of context on distinct facial expressions of disgust. *Emotion*, *19*(2), 365–370. https://doi.org/10.1037/emo0000445

Rice, K., & Redcay, E. (2015). Spontaneous mentalizing captures variability in the cortical thickness of social brain regions. *Social Cognitive and Affective Neuroscience*, *10*(3), 327–334. https://doi.org/10.1093/scan/nsu081

Rice, K., Viscomi, B., Riggins, T., & Redcay, E. (2014). Amygdala volume linked to individual differences in mental state inference in early childhood and adulthood. *Developmental cognitive neuroscience*, *8*, 153–163. https://doi.org/10.1016/j.dcn.2013.09.003

Righart, R., & de Gelder, B. (2008a). Rapid influence of emotional scenes on encoding of facial expressions: An ERP study. *Social Cognitive and Affective Neuroscience*, *3*(3), 270–278. https://doi.org/10.1093/scan/nsn021

Righart, R., & Gelder, B. de. (2008b). Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(3), 264–272. https://doi.org/10.3758/CABN.8.3.264

Ronfard, S., & Harris, P. L. (2014). When will Little Red Riding Hood become scared? Childrens attribution of mental states to a story character. *Developmental Psychology*, *50*(1), 283–292. https://doi.org/10.1037/a0032970

Ruckmick, C. A. (1921). A preliminary study of the emotions. *Psychological Monographs*, *30*(3), 30–35. https://doi.org/10.1037/h0093142

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 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. *Psychological Bulletin*, *115*(1), 102–141. https://doi.org/10.1037/0033-2909.115.1.102

Russell, J. A. (2016). A Sceptical Look at Faces as Emotion Signals. In C. Abell & J. Smith (Eds.), *The Expression of Emotion: Philosophical, Psychological and Legal Perspectives* (pp. 157–172). Cambridge University Press. https://doi.org/10.1017/CBO9781316275672.008

Russell, J. A., & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology: General*, *116*(3), 223–237. https://doi.org/10.1037/0096-3445.116.3.223

Russell, J. A., Suzuki, N., & Ishida, N. (1993). Canadian, Greek, and Japanese freely produced emotion labels for facial expressions. *Motivation and emotion*, *17*(4), 337–351. https://doi.org/10.1007/BF00992324

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92. https://doi.org/10.1177/1043463195007001004

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, *9*(4), 174–179. https://doi.org/10.1016/j.tics.2005.01.012

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15–21. https://doi.org/10.1016/j.copsyc.2017.04.019

Scherer, K. R. (2005). Appraisal Theory. In *Handbook of Cognition and Emotion* (pp. 637–663). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch30

Scherer, K. R., & Meuleman, B. (2013). Human Emotion Experiences Can Be Predicted on Theoretical Grounds: Evidence from Verbal Labeling. *PLoS ONE*, *8*(3), e58166–8. https://doi.org/10.1371/journal.pone.0058166

Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proc. ACM Hum.-Comput. Interact.*, *5*. https://doi.org/10.1145/3476058

Schlegel, K., Grandjean, D., & Scherer, K. R. (2012). Emotion recognition: Unidimensional ability or a set of modality- and emotion-specific skills? *Personality and Individual Differences*, *53*(1), 16–21. https://doi.org/10.1016/j.paid.2012.01.026

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., & Viding, E. (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, *7*(1), 53–63. https://doi.org/10.1093/scan/nsr023

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Cosmides, L., & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110–128. https://doi.org/10.1016/j.cognition.2017.06.002

Sen, T., Hasan, M. K., Tran, M., Yang, Y., & Hoque, M. E. (2018). Say CHEESE: Common human emotional expression set encoder and its application to analyze deceptive communication. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 357–364. https://doi.org/10.1109/FG.2018.00058

Shariff, A. F., & Tracy, J. L. (2011). What Are Emotion Expressions For? *Current Directions in Psychological Science*, *20*(6), 395–399. https://doi.org/10.1177/0963721411424739

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J., & Ullman, T. (2021). AGENT: A benchmark for core psychological reasoning. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 9614–9625, Vol. 139). PMLR.

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9*(6), 838–846. https://doi.org/10.1037/a0017810

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of neuroscience*, *34*(48), 15997–16008. https://doi.org/10.1523/JNEUROSCI.1676-14.2014

Skerry, A. 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

Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216. https://doi.org/10.1016/j.cognition.2013.11.002

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc.

Sowden, S., Schuster, B. A., Keating, C. T., Fraser, D. S., & Cook, J. L. (2021). The role of movement kinematics in facial emotion expression production and recognition. *Emotion*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/emo0000835

Srinivasan, R., Golomb, J. D., & Martinez, A. M. (2016). A neural basis of facial action recognition in humans. *Journal of Neuroscience*, *36*(16), 4434–4442. https://doi.org/10.1523/JNEUROSCI.1704-15.2016

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 0082. https://doi.org/10.1038/s41562-017-0082

Stojnic, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2022, June 11). Commonsense Psychology in Human Infants and Machines. https://doi.org/10.31234/osf.io/j3zs8

Sznycer, D. (2019). Forms and Functions of the Self-Conscious Emotions. *Trends in Cognitive Sciences*, *23*(2), 143–157. https://doi.org/10.1016/j.tics.2018.11.007

Sznycer, D., Sell, A., & Lieberman, D. (2021). Forms and Functions of the Social Emotions. *Current Directions in Psychological Science*, *30*(4), 292–299. https://doi.org/10.1177/09637214211007451

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual review of psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., & Porter, J. M. (2014). Misleading First Impressions. *Psychological Science*, *25*(7), 1404–1417. https://doi.org/10.1177/0956797614532474

Tracy, J. L. (2014). An Evolutionary Approach to Understanding Distinct Emotions. *Emotion Review*, *6*(4), 308–312. https://doi.org/10.1177/1754073914534478

Tracy, J. L., & Robins, R. W. (2004). Show Your Pride: Evidence for a Discrete Emotion Expression. *Psychological Science*, *15*(3), 194–197. https://doi.org/10.1111/j.0956-7976.2004.01503008.x

Tsividis, P. A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., Gershman, S. J., & Tenenbaum, J. B. (2021, July 26). *Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning*.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

van den Assem, M. J., van Dolder, D., & Thaler, R. H. (2012). Split or Steal? Cooperative Behavior When the Stakes Are Large. *Management Science*, *58*(1), 2–20. https://doi.org/10.1287/mnsc.1110.1413

Van Kleef, G. A. (2010). The Emerging View of Emotion as Social Information. *Social and Personality Psychology Compass*, *4*(5), 331–343. https://doi.org/10.1111/j.1751-9004.2010.00262.x

van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-09161-6

Wallbott, H. G. (1988). In and out of context: Influences of facial expression and context information on emotion attributions. *British Journal of Social Psychology*, *27*(4), 357–369. https://doi.org/10.1111/j.2044-8309.1988.tb00837.x

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. *The Journal of neuroscience*, *34*(20), 6813–6821. https://doi.org/10.1523/JNEUROSCI.4478-13.2014

The weeping Oscar Pistorius and a final question: Has it all been an act? (2014).

Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S., Bien, C. G., & Kissler, J. (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *CORTEX*, *69*(100), 131–140. https://doi.org/10.1016/j.cortex.2015.05.003

Weimer, A. A., Sallquist, J., & Bolnick, R. R. (2012). Young Children's Emotion Comprehension and Theory of Mind Understanding. *Early Education & Development*, *23*(3), 280–301. https://doi.org/10.1080/10409289.2010.517694

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wenzler, S., Levine, S., van Dick, R., Oertel-Knöchel, V., & Aviezer, H. (2016). Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion*, *16*(6), 807–814. https://doi.org/10.1037/emo0000185

Widen, S. C., & Russell, J. A. (2015). Do Dynamic Facial Expressions Convey Emotions to Children Better Than Do Static Ones? *Journal of Cognition and Development*, *16*(5), 802–811. https://doi.org/10.1080/15248372.2014.916295

Widen, S. C., & Russell, J. A. (2010). Differentiation in preschooler's categories of emotion. *Emotion*, *10*(5), 651–661. https://doi.org/10.1037/a0019005

Wieser, M. J., Gerdes, A. B., Büngel, I., Schwarz, K. A., Mühlberger, A., & Pauli, P. (2014). Not so harmless anymore: How context impacts the perception and electrocortical processing of neutral faces. *NeuroImage*, *92*, 74–82. https://doi.org/10.1016/j.neuroimage.2014.01.022

Williams, W. C., Leong, Y. C., Collier, E. A., Nook, E. C., Son, J.-Y., & Zaki, J. (2021, May). *Communicating emotion through facial expressions: Social consequences and neural correlates* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/4zpfk

Witkower, Z., Tracy, J. L., Hill, A., & Koster, J. (2020). Beyond face value: Evidence for the universality of bodily expressions of emotion. *Affective Science*.

Witkower, Z., & Tracy, J. L. (2019). Bodily Communication of Emotion: Evidence for Extrafacial Behavioral Expressions and Available Coding Systems. *Emotion Review*, *11*(2), 184–193. https://doi.org/10.1177/1754073917749880

Wondra, J. D., & Ellsworth, P. C. (2015). An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review*, *122*(3), 411–428. https://doi.org/10.1037/a0039252

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational Inference of Beliefs and Desires From Emotional Expressions. *Cognitive Science*, *42*(3), 850–884. https://doi.org/10.1111/cogs.12548

Wu, Y., Muentener, P., & Schulz, L. E. (2017). One- to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proceedings of the National Academy of Sciences*, *114*(45), 11896–11901. https://doi.org/10.1073/pnas.1707715114

Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as Information in Early Social Learning. *Current Directions in Psychological Science*, *30*(6), 468–475. https://doi.org/10.1177/09637214211040779

Xu, T., White, J., Kalkan, S., & Gunes, H. (2020, August 21). Investigating Bias and Fairness in Facial Expression Recognition.

Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 435–442. https://doi.org/10.1145/2818346.2830595

Zaki, J. (2013). Cue Integration: A Common Framework for Social Cognition and Physical Perception. *Perspectives on Psychological Science*, *8*(3), 296–312. https://doi.org/10.1177/1745691613475454

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion*, *13*(5), 803–810. https://doi.org/10.1037/a0033839

Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022, August 4). Solving the Baby Intuitions Benchmark with

a Hierarchically Bayesian Theory of Mind. https://doi.org/10.48550/ARXIV.2208.02914