Social Emotions and Psychological Games

Luke J. Chang[1*]
Alec Smith[2*]

[1] Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH 03755
luke.j.chang@dartmouth.edu

[2] Department of Economics (MC 0316)
Virginia Tech
880 West Campus Drive
Blacksburg, VA 24061
alecsmith@vt.edu

*Corresponding author. Both authors contributed equally to this work, and author order is alphabetical.

Word Count: 3366

Abstract

Emotions arise from cognitive appraisals and organize adaptive behavioral responses. The appraisals associated with social emotions such as guilt and anger can be mathematically modeled with utility functions that depend on both material and psychological payoffs. These functions can predict each agent's decisions and payoffs in strategic interactions. Guilt arises from the belief that an agent has disappointed a relationship partner and motivates reparative actions, while anger arises from the frustration of a goal being unexpectedly blocked and motivates aggressive actions. These psychological payoffs not only enable cooperation, but also appear to be associated with neural activations consistent with negative affective states. We believe integrating appraisal theory with game theoretic modeling can improve our ability to study emotions and predict behavior in social interactions.

We spend much of our waking day engaged in collaborative social exchanges. Social emotions, in particular, are central in ensuring the success of these interactions [1]. For example, consider a scientific collaboration in which A shares data with B to perform a specialized analytic technique with the goal of publishing the results together. B would likely feel guilty if he shirked his responsibility and in response A would be angry that she wasted her time trusting B. Thus, social emotions play a critical role in ensuring a successful exchange. These emotions can be modeled using utility functions that incorporate both material and psychological payoffs, and their effect on behavior can be mathematically described using game theory. In this article we focus on two social emotions, guilt and anger, and demonstrate how game-theoretic models of these emotions capture important aspects of social behavior.

Psychological models of emotion

Emotions are psychological states comprised of multiple interrelated processes such as cognitive appraisals, physiological responses, behavioral action tendencies, and the phenomenological experience of feelings. Though there are many different perspectives on emotion ranging from categorical models of discrete emotions [2,3], multi-dimensional factor models [4,5], and psychological constructionist models [6,7], none have been more amenable to computational modeling than the cognitive framework of appraisal theory [8-11]. Appraisal theory defines emotions as adaptive responses that are elicited based on how an agent evaluates its situation (e.g., novelty, valence, threat, contamination, social norms, etc.) [9,12-15]. Appraisals are typically directly related to the motivational goals of the agent (e.g., basic needs, safety, cultural values, beliefs) and occur in response to both external stimuli and also to internally generated thoughts, e.g.

when the agent is imagining the future or remembering the past. People continually interpret their environment with respect to their motivational goals and these evaluations or appraisals give rise to different feeling states that evolve as information changes [9,14]. Appraisals are thus cognitive antecedents to the experience of the emotion, though it remains an open question whether emotions are a consequence of appraisals or if the appraisal itself constitutes the emotional experience [16].

In our view, appraisals precede emotions, which in turn prepare the agent to make adaptive responses via action tendencies [17]. Action readiness is the state of translating feelings and goals into behavioral actions. These actions could be as simple as approaching or avoiding a stimulus [18], or could take the form of embodied action preparations [19,20]. Whereas appraisals describe the inputs of the emotional experience, action tendencies delineate the behavioral outputs. This input-output view of emotion provides a structure that can be translated into mathematical models.

In this paper we focus on guilt and anger, two emotions that arise from social interactions and which can be described in terms of cognitive appraisals and action tendencies [21].[1] For example, guilt arises from the appraisal that one has failed to live up to the expectations of a relationship partner [24] and motivates reparative action tendencies [25-27]. Anger, in contrast, arises from the appraisal that progress towards a goal is blocked, or a social/moral norm has been transgressed [28,29], and motivates punishment and revenge action tendencies [16,30]. Using a theoretical approach known as psychological game theory [31,32] the appraisals associated with these emotions

[1] All emotions can be considered "social" to some degree as they involve communicating internal states [3,22,23].

may be captured as the changes in an agent's expected payoff following a new event or outcome. These belief dependent appraisals can then be directly incorporated into the agent's utility function as psychological payoffs (i.e., subjective feelings) to capture the action tendencies associated with emotions.

Game theoretic models of emotion

Game theory is a set of mathematical tools for modeling interactive decision-making. These include mathematical descriptions of the strategies available to the players and of the payoffs (or utilities) resulting from those strategies. Additional details may include the sequence of play, the actions available to each player at each stage of the game, and the information available to each player. Players' beliefs are represented via probability distributions over actions, states, or other players' beliefs [33]. When combined with solution concepts such as Nash equilibrium [34], subgame perfect equilibrium [35], or sequential equilibrium [36], the formal structure of game theory provides predictions about how the game will be played and the payoffs to each player.

Early game theoretic models and applications assumed that agents behaved selfishly in maximizing their material self-interest: that is, each player's utility function depended only upon his own payoff. These models of purely self-interested individuals perform poorly in predicting social behavior. For example, they predict unrealistically low levels of voter turnout and charitable donation [37,38]. In addition, countless laboratory experiments have shown that people often behave unselfishly (e.g., sharing resources, punishing malefactors) [39]. A number of different theoretical models attempt to capture this other-regarding behavior by modifying the standard selfish utility function to include

concerns for social factors such as inequality [40,41], social welfare [42], fairness & reciprocity [43-45], or social image [46]. These models of social preferences enable pairs or groups of individuals to obtain outcomes that purely self-interested individuals cannot [47].

In this paper, we explore models of other-regarding preferences that directly incorporate emotional rewards into the payoffs players receive in the game. These models allow players' payoffs and psychological states to depend upon their beliefs, using tools from psychological game theory [31,32]. As noted by Geanakoplos et al. [31], and consistent with the appraisal theory approach to emotions, "A player's emotional reactions cannot in general be independent of his expectations and of his interpretation of what he learns in a play of a game." The psychological games approach thus requires the modeler to make precise assumptions about the appraisal triggers of emotions and the resulting consequences of those emotions for behavior. In addition, psychological game theory is well suited for modeling the theory-of-mind reasoning that is often associated with social emotions [48,49]. Both the appraisals and the action tendencies associated with guilt and anger, for example, can be modeled by adding a psychological payoff term to the standard material payoff. This approach highlights that agents face tradeoffs between psychological and material payoffs, so that emotions need not always result in a pre-programmed action.

While we focus on the behavioral predictions of the models, we believe that "psychological payoffs" are real and can be validated by their neural and physiological correlates. We therefore also report on the results of fMRI and other studies that seek to identify physiological data that corresponds to certain emotions. Ultimately, the models

we describe will either be falsified or supported via a combination of behavioral and physiological data.

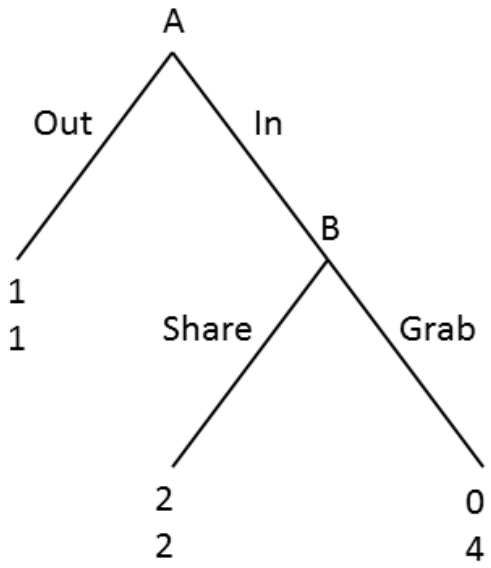
Guilt

Battigalli & Dufwenberg (2007) [50] develop a model whereby a player feels guilty to the extent his actions cause a co-player to receive less than he expected (see also [51-53]). Player A's strategy is denoted by s_A and his material payoff by π_A. A given history of the game is denoted by h. Player A's guilt towards player B is determined by the function $G_{AB} = max(E_B[\pi_B|h_0] - \pi_B, 0)$, where $E_B[\pi_B|h_0]$ represents B's expected payoff, calculated at the initial history (the start) of the game with respect to B's first-order beliefs and his strategy.[2] However, player A does not know what payoff player B initially expected. So player A's expected utility $E_A^1[U_A]$ is a combination of material and psychological payoffs and is calculated with respect to his second-order beliefs: $\mathrm{E}_A^2[U_A(s_A)|h] = \mathrm{E}_A^2[\pi_A(s_A) - \theta_A G_{AB}(s_A)|h]$, where θ_A is a parameter reflecting Player A's sensitivity to guilt. Battigalli & Dufwenberg [50] refer to this model as "simple guilt."[3].

[2] For ease of exposition, we often suppress the notation showing the dependence of each players' payoffs on strategies. More formally each players' material payoff is a function of the strategies of all coplayers, so that e.g. $\pi_A = \pi_A(s_A, s_B)$ in two player games.

[3] They also develop "guilt from blame," where A feels guilty for letting down B only when A believes that B believes that A caused B to get less than he expected. We refer the interested reader to Battigalli & Dufwenberg (2007) [50] for the formal model.

Figure 1. Trust game

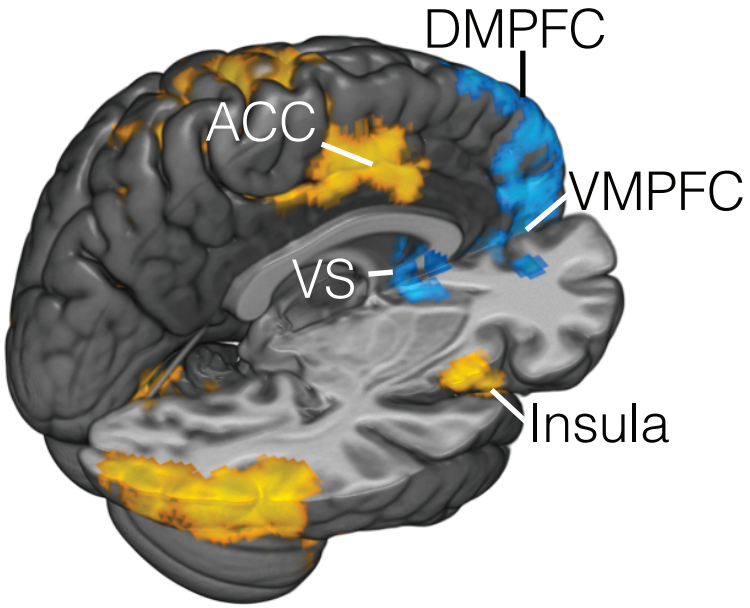


The Trust Game reported in [54]. Player A can choose "In" or "Out". If Player A chooses out, the players each get 1. If Player A chooses "In", the amount is multiplied by 4 and Player B can choose to "Keep" all of the money or "Share" half with Player A.

To illustrate, consider a simple trust game as shown in Figure 1 from [54]. Player A can either choose "In" or "Out", while Player B can choose between "Share" and "Grab." If both players are "selfish" in the sense that they maximize their material payoffs, then the unique subgame perfect equilibrium is for Player A to choose "Out." This inefficient outcome results because a selfish Player B will choose to "Grab" if given the opportunity. However, the behavioral prediction changes when Player B is averse to guilt. If Player B believes that Player A expects Player B to choose "Share," then a Player B who is sufficiently sensitive to guilt will choose "Share" in order to avoid the guilt that would result from selecting "Grab." This example shows how guilt can encourage trust and cooperation in social interactions.

Laboratory studies have found behavioral evidence supporting the predictions of guilt aversion. Dufwenberg & Gneezy [55] found that the amount of money that Player B returns in a trust game is directly proportional to the amount that they believe Player A expects them to return. Charness & Dufwenberg [52] observed that allowing players to send messages before playing a Trust Game resulted in large increases in the cooperation subgames compared to not sending messages [52]. Reuben et al. [56] found that participants were more likely to reciprocate when they believed that their partner had higher expectations of them cooperating. Recent experimental work by Khalmetski et. al (2015) [57] and Ederer & Stremitzer (2015) [58] also provides evidence consistent with guilt aversion. Together these results provide behavioral evidence that belief-dependent guilt enables cooperation in a manner consistent with the predictions of guilt aversion theory [59].

Figure 2. Neural Responses to Guilt aversion



This figure displays the results from [62], in which Player B decides how much money to return to Player A. Areas in orange are associated with decisions that minimize anticipated guilt and include the anterior cingulate cortex (ACC) and insula. Areas in blue are associated with decisions to maximize material payoffs and include the ventral striatum (VS), ventromedial prefrontal cortex (VMPFC), and dorsomedial PFC (DMPFC).

These studies demonstrate that players exhibit concern for their co-players' expectations in a manner consistent with guilt aversion. However, it remains an open question if this psychological payoff resembles a negative affective state. One way to study whether participants experience guilt aversion is to examine their neural activity while they make decisions. Chang et al. [60] scanned participants while they played the role of Player B in a multi-round single-shot trust game using functional magnetic resonance imaging (fMRI). Importantly, Chang et al. elicited Player B's second-order beliefs regarding their co-player's expectations about B's own behavior. This allowed the authors to compare

trials in which participants chose a strategy that minimized guilt to trials in which they chose a strategy that was more self-interested. This contrast revealed neural activity in two distinct brain systems associated with value (Figure 2). When participants behaved as-if guilt averse, they had increased activity in the insula, anterior cingulate cortex (ACC) dorsolateral prefrontal cortex (DLPFC) and temporoparietal junction (TPJ), a network that is thought to be involved in processing negative affect, salience, cognitive control, and theory of mind. When participants behaved in accordance with maximizing financial self-interest, they had increased activation in the ventromedial prefrontal cortex (VMPFC), ventral striatum, and dorsomedial prefrontal cortex (DMPFC), regions consistently involved in reward processing and mentalizing. Furthermore, post-experimental ratings of counterfactual guilt revealed that participants reported that they would have experienced more guilt had they returned a lesser amount of money and that this correlated with the magnitude of the response in their insula when they behaved as-if guilt averse. These results provide neural evidence consistent with guilt aversion theory, in which players have competing motivations to maximize material payoffs and to minimize the aversive psychological payoffs from disappointing a relationship partner.

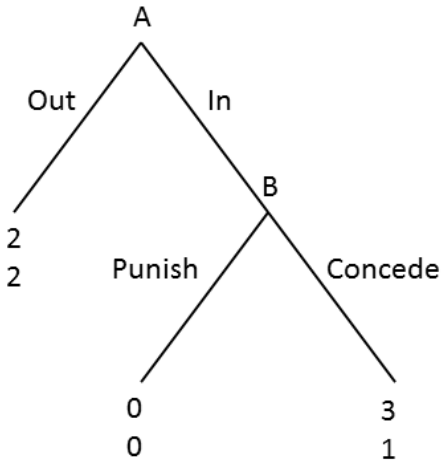
Anger

Battigalli, Dufwenberg & Smith [61] connect anger with frustration, modeling anger as a function of the blame player A places on player B for his frustration. Frustration is defined as the difference between the best outcome player A can still receive in the game and the material payoff player A had initially expected. That is, A's frustration after history h is $F_A(h) = max\left\{E_A^1[\pi_A|h_0] - \max_{s_A|h} E_A^1[\pi_A(s_A)|h], 0\right\}$. In this formulation, after

history h, player A compares the best possible payoff he can still get $(\max_{s_A|h} E_A^1[\pi_A(s_A|h])$ to the payoff he expected at the start of the game ($E_A^1[\pi_A|h_0]$). If the difference is positive, A is frustrated. Anger is modeled as the decreased weight that player A places on player B's material payoff if he is frustrated, so that a player who is prone to anger maximizes the expected utility function $E_A^1[U_A(s_A)|h] = E_A^1[\pi_A(s_A) - \theta_A F_A(h)\pi_B]$, where θ_A is A's personal sensitivity to anger. With this formulation, referred to as "simple anger", even frustrations not attributed to player B can result in player A choosing actions that harm player B.[4]

[4] Battigalli et al. [61] also develop models of anger where A's anger towards B is modulated by cognitive appraisals of blame. In "anger from blaming behavior" player A only directs anger towards active players. In "anger from blaming intentions" A's anger is restricted towards players that A believes intended to give him a low payoff.

Figure 3. Ultimatum Minigame



The Ultimatum Minigame reported in [63]. Player A is endowed with \$4 and can choose to go "Out" and split the money evenly with Player B, or can choose "In" which amounts to a demand for a larger share of the pie. If Player A chooses "In", Player B then decides whether to "Punish" Player A, in which neither player receives a payoff, or to "Concede" and allow Player B to receive a larger payoff.

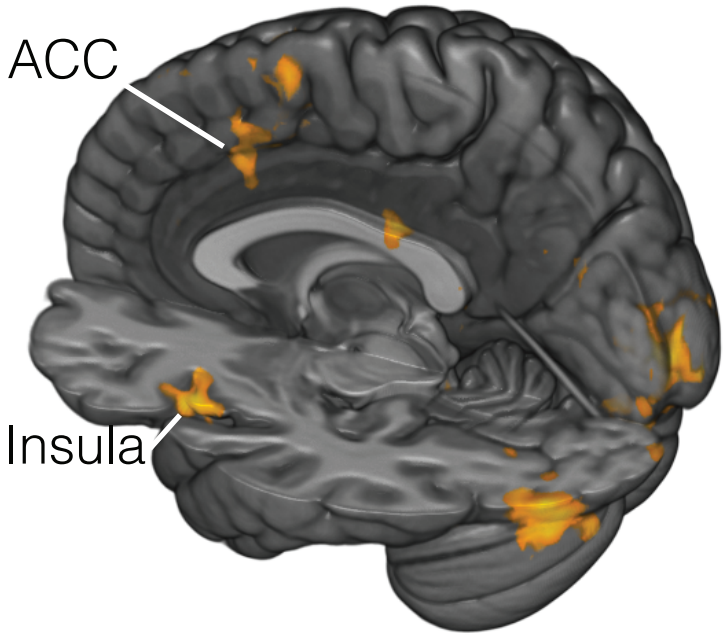
Consider the ultimatum minigame from [61] shown in Figure 3. Player A can choose "In" or "Out". If Player A chooses "In," Player B then decides whether to "Punish" Player A for not selecting a more equitable division of money or "Concede" in which case Player B will receive less money than Player A. In this game, if both players are selfish, there is a unique subgame perfect equilibrium where Player A's strategy is to choose "In" and Player B's is to choose "Concede." Because Player B prefers higher to lower material payoffs he chooses "Concede." Knowing this, Player A will choose "In" to get the larger payoff. However, the solution concept changes if Player B experiences anger. If Player B receives a sufficiently large psychological payoff from anger ($\theta_B \gg 0$), he will choose to "Punish" if Player A chooses "In." Knowing this, Player A will choose "Out," so both

players will end up with equal payoffs. This example illustrates how the threat of punishment can ensure a cooperative outcome.[5]

A number of behavioral studies have found support for this conceptualization of frustrated anger, which arises following worse than expected outcomes [61,63] Laboratory studies have found that participants reported feeling more anger when another player strongly deviated from the normative contribution in a public goods game, which led to incurring a cost to punish these players [64]. Receiving lower offers than initially expected in the Ultimatum Game results in feelings of anger [30] and increased rejection rates [65,66]. Furthermore, in "power-to-take" experiments, differences between expectations and outcomes results in anger, which in turn drives costly punishment [67,68]. Further support for the frustrated anger model can also be found outside the laboratory. Incidences of domestic violence measured from police reports appear to increase following unexpected losses by local professional football teams [62]. Together, these studies provide compelling support for the notion that anger arises when one experiences a worse outcome than initially expected.

[5] Battigalli et al. [61] also illustrate the potentially destructive effects of frustration-driven anger, where frustrated agents become generally hostile and aggressive (e.g. Card & Dahl, 2011 [62]).

Figure 4. Neural Responses to Anger



This figure displays the results from [73], in which Player B receives an offer from Player A in an Ultimatum Game. Areas in orange are parametrically associated with offers that deviate from Player A's expectations about the types of offers they expect to encounter. These regions include the anterior cingulate cortex (ACC) and insula.

Further evidence for the expectation-based anger model can be found in neuroimaging studies [69,70]. Chang & Sanfey (2013) [71] elicited Player B's beliefs about the most frequent offers they expected to encounter in a single-shot multi-round Ultimatum Game and found that players were more likely to reject offers that deviated from their initial expectations. Importantly, the magnitude of the deviation from expectations correlated with activity in the dorsal ACC (dACC) and the anterior insula, a network reliably associated with error-monitoring and emotion (Figure 4). Another study manipulated expectations by varying the distribution of offers that the players encountered in the game such that they were initially drawn from either a high or low Gaussian distribution and then examined the effect of these expectations on offers drawn from an intermediate

distribution [72]. The authors found that players who were conditioned on the high distribution relative to the low distribution were more likely to reject intermediate offers in the game and report feeling higher levels of negative affect. Similar to Chang et al. [71] the magnitude of trial-to-trial deviations from expectations (represented as an ideal bayesian observer) correlated with the anterior insula, dACC, and ventral striatum. Importantly, this signal also corresponded to trial-to-trial negative affective ratings. A recent follow-up study found that lesions to the insula, but not the vmPFC disrupted the ability to adapt from this error signal [73]. Together these results suggest that the frustrated anger model is accurately capturing the appraisal computation and that the consequent feelings of anger, in turn, motivate decisions to punish a transgressor.

Conclusions and suggestions for future work

In this paper we have discussed how the social emotions guilt and anger can be conceptualized in terms of appraisal theory and formally modeled as negative psychological payoffs based on players' beliefs using game theoretic models. Importantly, models that incorporate social emotions theoretically predict enhanced cooperation in both trust and bargaining games. We reviewed behavioral evidence illustrating that the theoretical predictions of these models better describe behavior than models that solely consider material payoffs. Furthermore, we reviewed several neuroimaging studies, which show that the appraisal computations associated with these psychological payoffs have biological substrates that are consistent with the negative emotional feelings of guilt and anger.

We note several interpretive caveats to our argument. First, we have framed the use of neuroimaging as a method to validate the appraisal processes that are predicted by our psychological game-theoretic models. For example, we take the ACC and insula activations that correlate with our model as evidence supporting the expectation violation appraisal process predicted by our anger model. However, the insula and ACC are activated in about 40% of all neuroimaging studies [74], which means we are currently unable to make strong inferences about a specific psychological state upon observing their activation, an inferential fallacy referred to as reverse-inference [75]. We (and others) are actively developing techniques that allow us to make stronger inferences about affective states predicted from brain activation using supervised machine learning techniques [76-78] that will improve our ability to make reverse inferences. We believe that validating the psychological components of our models is an important endeavor in the model building process that complements more traditional behavioral validations.

Second, the game-theoretic models we discuss in this paper have some limitations. The models have two components: a utility function that depends upon both material and psychological components, and a solution concept (such as sequential equilibrium). The models generate emotions via the comparison of expected payoffs with outcomes or updated expected payoffs. These expectations are modeled explicitly as mathematical expected value calculations, but more theoretical work is needed to capture additional factors (e.g. attention) that may influence the computation of these reference points or of expected payoffs in general (see e.g. Gneezy & Imas, [79]). In addition, the equilibrium solution concepts employed assume that beliefs are (on average) correct. In practice, behavior and/or beliefs may be out-of-equilibrium, so that studying the effect of belief-

dependent emotions often requires the measurement of beliefs in order to make accurate predictions about behavior.

Though we have focused on (simple) guilt and anger in this review, there are many promising areas for future work. In particular, there are numerous other social emotions that could be modeled using tools from psychological game theory, such as shame [80]. We also believe our models have a direct analogue to positive emotions that result from receiving better-than-expected outcomes. For example, a player might feel generous if he shared more than his partner expected such as a "surprising gift" [57] which may be related to charitable giving and warm-glow altruism [81-83]. Also, players may feel gratitude upon receiving more than they expected, which might motivate indebtedness action tendencies [84] and positive reciprocity. Another promising direction for future work might be to explore the broader societal implications of social emotions, for example in the form of multi-agent simulations [85] or macro-level structural models. More generally, integrating appraisal theory with game theory provides an important conceptual advance for predicting when emotions will be experienced and how they will impact social behavior. This work could have profound implications for institutional policies that typically only consider self-interest [1,86,87].

Annotated Bibliography

• of special interest
•• of outstanding interest

1. Smith A: *The theory of moral sentiments*. Indianapolis: Liberty Fund; 1984 (1759).
2. Izard CE, Ackerman BP: Motivational, organizational, and regulatory functions of discrete emotions. *Handbook of emotions* 2000, 2:253-264.
3. Ekman P: Facial expression and emotion. *American psychologist* 1993, 48:384.
4. Schachter S, Singer J: Cognitive, social, and physiological determinants of emotional state. *Psychological review* 1962, 69:379.
5. Russell JA: A circumplex model of affect. *Journal of personality and social psychology* 1980, 39:1161.
6. Barrett LF: Are Emotions Natural Kinds? *Perspectives on Psychological Science* 2006, 1:28-58.
7. Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF: The brain basis of emotion: a meta-analytic review. *Behav Brain Sci* 2012, 35:121-143.
8. Scherer KR: Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2009, 364:3459-3474.
•9. Ortony A, Clore GL, Collins G: The cognitive structure of emotions. Edited by: Cambridge University Press; 1988.

    A classic book outlining the cognitive structure of emotions. Provides one of the first computational appraisal models of emotion.

10. Armony JL, Servan-Schreiber D, Cohen JD, LeDoux JE: Computational modeling of emotion: Explorations through the anatomy and physiology of fear conditioning. *Trends in Cognitive Sciences* 1997, 1:28-34.
11. Marsella S, Gratch J, Petta P: Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual* 2010:21-46.
12. Smith CA, Ellsworth PC: Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology* 1985, 48:813.
13. Scherer KR: Appraisal Theory. In *Handbook of cognition and emotion*. Edited by; 1999:637-663.
•14. Ellsworth PC, Scherer KR: Appraisal processes in emotion. Handbook of affective sciences 2003, 572:V595.

    A comprehensive overview of the history and evidence for emotion appraisal theory.

15. Oatley K, Johnson-Laird P: Cognitive approaches to emotions. *Trends in cognitive sciences* 2014, 18:134-140.
••16. Frijda NH: The place of appraisal in emotion. Cognition & Emotion 1993, 7:357-387.

    A classic paper outlining the appraisal processes involved in guilt and anger.

17. Frijda NH, Kuipers P, Ter Schure E: Relations among emotion, appraisal, and emotional action readiness. *Journal of personality and social psychology* 1989, 57:212.
18. Davidson RJ, Irwin W: The functional neuroanatomy of emotion and affective style. *Trends Cogn Sci* 1999, 3:11-21.
19. Niedenthal PM: Embodying emotion. *Science* 2007, 316:1002-1005.
20. Niedenthal PM, Barsalou LW, Winkielman P, Krauth-Gruber S, Ric F: Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review* 2005, 9:184-211.
•21. Haidt J: The moral emotions. In Handbook of affective sciences. Edited by Davidson RJ, Scherer KR, Goldsmith HH: Oxford University Press; 2003:852-870.

An accessible review of the appraisal and action tendencies associated with social emotions.

22. Parkinson B: Emotions are social. *British journal of psychology* 1996, 87:663-683.
23. Darwin C: *The expression of the emotions in man and animals*: Oxford University Press; 2002.
••24. Baumeister RF, Stillwell AM, Heatherton TF: Guilt: an interpersonal approach. Psychol Bull 1994, 115:243-267.

A comprehensive overview of psychological research on guilt. Proposes that guilt results from the failure to meet a relationship partner's expectations.

25. Regan DT, Williams M, Sparling S: Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology* 1972, 24:42.
26. Tangney JP, Miller RS, Flicker L, Barlow DH: Are shame, guilt, and embarrassment distinct emotions? *J Pers Soc Psychol* 1996, 70:1256-1269.
27. Carlsmith JM, Gross A: Some effects of guilt on compliance. *Journal of Personality and Social Psychology* 1969, 59:538-549.
•28. Dollard J, Miller NE, Doob LW, Mowrer OH, Sears RR: Frustration and aggression. 1939.

The original source of the frustration-aggression hypothesis.

29. Berkowitz L: Frustration-Aggression Hypothesis: Examination and Reformulation. *Psychological Bulletin* 1989, 106:59-73.
30. Pillutla MM, Murnighan JK: Unfairness, anger, and spite: Emotional rejection of ultimatum offers. *Organizational Behavior and Human Decision Processes* 1996, 68:208.
•31. Geanakoplos J, Pearce D, Stacchetti E: Psychological games and sequential rationality. *Games and Economic Behavior* 1989, 1:60-79.

This paper provides a framework for incorporating belief-dependent motivations into game theoretic models.

••32. Battigalli P, Dufwenberg M: Dynamic psychological games. *Journal of Economic Theory* 2009, 144:1-35.

This paper builds on research in dynamic interactive epistemology to provide a more general framework for psychological games. The framework allows for payoffs to depend upon updated higher-order beliefs, others' beliefs, and plans of action, making it possible to capture dynamic psychological effects.

33. Battigalli P, Siniscalchi M: Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games. *Journal of Economic Theory* 1999, 88:188-230.
34. Nash JF: Equilibrium points in n-person games. *Proceedings of the national academy of sciences* 1950, 36:48-49.
35. Selten R: Reexamination of the perfectness concept for equilibrium points in extensive games. *International journal of game theory* 1975, 4:25-55.
36. Kreps DM, Wilson R: Sequential equilibria. *Econometrica: Journal of the Econometric Society* 1982:863-894.
37. Kahneman D, Knetsch J, Thaler RH: Fairness and the assumptions of economics. *Journal of Business* 1986, 59:S285-S300.
38. Rotemberg JJ: Models of Caring, or Acting as if One Cared, About the Welfare of Others. *Annu. Rev. Econ.* 2014, 6:129-154.
•39. Camerer CF: Behavioral Game Theory. New York: Russell Sage Foundation; 2003.

Chapter 2 reviews experimental evidence for other-regarding preferences and theoretical models developed in response.

40. Fehr E, Schmidt KM: A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 1999, 114:817-868.
41. Bolton GE, Ockenfels A: ERC: A theory of equity , reciprocity, and competition. *American Economic Review* 2000, 90:166-193.
42. Charness G, Rabin M: Understanding social preferences with simple tests. *Quarterly journal of Economics* 2002:817-869.
••43. Rabin M: Incorporating fairness into game theory and economics. *American Economic Review* 1993, 83:1281-1302.

This paper develops a model of intentions-based reciprocity in a psychological games framework. Players derive psychological payoffs from a reciprocal kindness function. The kindness function depends upon players' beliefs and beliefs about their co-players' beliefs.

•44. Dufwenberg M, Kirchsteiger G: A theory of sequential reciprocity. *Games Econ Behav* 2004, 47:268-298.

A model of intentions-based reciprocity for sequential games.

45. Falk A, Fischbacher U: A theory of reciprocity. *Games Econ Behav* 2006, 54:293-315.

46. Andreoni J, Bernheim BD: Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* 2009, 77:1607-1636.
•47. Fehr E, Camerer CF: Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn Sci* 2007, 11:419-427.

An accessible overview of economic models of social preferences and their neural substrates.

48. Frith CD, Frith U: The neural basis of mentalizing. *Neuron* 2006, 50:531-534.
49. Saxe R: Uniquely human social cognition. *Current opinion in neurobiology* 2006, 16:235-239.
••50. Battigalli P, Dufwenberg M: Guilt in Games. *The American Economic Review* 2007, 97:170-176.

Develops two psychological games models of guilt-aversion. In simple guilt, Player A feels guilty when he believes he "let down" a co-player relative to expectations. In guilt from blame, Player A feels guilty when he believes that a Player B believes that A intended to let B down.

•51. Dufwenberg M: Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization* 2002, 48:57-69.

This paper shows how guilt aversion promotes cooperation and encourages investment.

52. Charness G, Dufwenberg M: Promises and Partnership. *Econometrica* 2006, 74:1579-1601.
53. Charness G, Dufwenberg M: Participation. *The American Economic Review* 2011, 101:1211-1237.
54. Dufwenberg M: Marital investments, time consistency and emotions. *Journal of Economic Behavior and Organization* 2002, 48:57-69.
55. Dufwenberg M, Gneezy U: Measuring beliefs in an experimental lost wallet game. *Games Econ Behav* 2000, 30:163-182.
56. Reuben E, Sapienza P, Zingales L: Is mistrust self-fulfilling? *Economics Letters* 2009, 104:89-91.
57. Khalmetski K, Ockenfels A, Werner P: Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory* 2015.
58. Ederer F, Stremitzer A: Promises and Expectations. *Cowles Foundation Discussion Paper No. 1931* 2015.
59. Kawagoe T, Narita Y: Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization* 2014, 102:1-9.
••60. Chang LJ, Smith A, Dufwenberg M, Sanfey AG: Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 2011, 70:560-572.

This study provided one of the first direct tests of the guilt-aversion theory using fMRI. Participants played a Trust Game and Player B was scanned while undergoing fMRI. Players' first and second order beliefs were elicited. Participants exhibited increased activation in the insula, ACC, DLPFC, and TPJ when they made decisions that minimized disappointing

their relationship partner and increased activation in the VMPFC and ventral striatum when they made decisions that maximized their material payoffs. This study provides evidence that negative psychological payoff associated with anticipated guilt is processed in regions of the brain associated with negative affect and error-monitoring.

••61. Battigalli P, Dufwenberg M, Smith A: Frustration and Anger in Games. *CESifo Working Paper Series No. 5258* 2015.

This paper develops a general model of frustration and anger in games. Worse-than-expected outcomes result in frustration, which lowers the weight that players place on their co-players material payoffs, and inceases the psychological payoff from aggressive behavior. The authors also develop two models that illustrate how blame can modulate anger, anger from blaming behavior and anger from blaming intentions.

62. Card D, Dahl GB: Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics* 2011, 126:103.
63. Dollard J, Doob L, Miller N, Mowrer O, Sears R: *Frustration and Aggression*. New Haven, CT: Yale University Press; 1939.
•64. Fehr E, Gachter S: Altruistic punishment in humans. *Nature* 2002, 415:137-140.

This study shows that players experience a negative affective response to free-riders in a public goods game and will incur a cost to punish them, which increases overall cooperation in the game.

65. Sanfey AG: Expectations and social decision-making: biasing effects of prior knowledge on Ultimatum responses. *Mind & Society* 2009, 8:93-107.
66. Schotter A, Sopher B: Advice and behavior in intergenerational ultimatum games: An experimental approach. *Games and Economic Behavior* 2007, 58:365 - 393.
67. Bosman R, Van Winden F: EMOTIONAL HAZARD IN A POWER‐TO‐TAKE EXPERIMENT. *The Economic Journal* 2002, 112:147-169.
68. Bosman R, Sutter M, van Winden F: The impact of real effort and emotions in the power-to-take game. *Journal of Economic Psychology* 2005, 26:407-429.
69. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD: The neural basis of economic decision-making in the Ultimatum Game. *Science* 2003, 300:1755-1758.
••70. Montague PR, Lohrenz T: To detect and correct: norm violations and their enforcement. *Neuron* 2007, 56:14-18.

This article proposes that both agents in a social interaction detect deviations from a shared social expectation and are motivated to comply and enforce adherence to the norm. These appraisals and action tendencies are consistent with guilt and anger.

••71. Chang LJ, Sanfey AG: Great expectations: neural computations underlying the use of social norms in decision-making. *Soc Cogn Affect Neurosci* 2013, 8:277-284.

Examines brain systems associated with appraial of descriptive norm violation in the ultimatum game. Trial-to-trial deviations in Player A's offers from Player B's expectation about the type of offers they expect to encounter are monotonically associated with increased activation in the insula and ACC. This appraisal is the antecedent of simple anger.

••72. Xiang T, Lohrenz T, Montague PR: Computational Substrates of Norms and Their Violations during Social Exchange. *J Neurosci* 2013, 33:1099-1108.

This study manipulates Player B's expectations about the normative offer in the ultimatum game. Trial-to-trial deviations in expectations correlate with activity in the insula, ACC, and ventral striatum. Importantly, this computational signal is also associated with participant's self-reported negative affect in response to offers.

73. Gu X, Wang X, Hula A, Wang S, Xu S, Lohrenz TM, Knight RT, Gao Z, Dayan P, Montague PR: Necessary, Yet Dissociable Contributions of the Insular and Ventromedial Prefrontal Cortices to Norm Adaptation: Computational and Lesion Evidence in Humans. *The Journal of Neuroscience* 2015, 35:467-473.
74. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD: Large-scale automated synthesis of human functional neuroimaging data. *Nature methods* 2011, 8:665-670.
75. Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, Nichols TE: Guidelines for reporting an fMRI study. *NeuroImage* 2008, 40:409-414.
76. Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD: A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLoS Biol* 2015, 13:e1002180.
77. Chang LJ, Yarkoni T, Khaw MW, Sanfey AG: Decoding the Role of the Insula in Human Cognition: Functional Parcellation and Large-Scale Reverse Inference. *Cerebral Cortex* 2013, 23:739-749.
78. Smith A, Bernheim BD, Camerer CF, Rangel A: Neural Activity Reveals Preferences without Choices. *American Economic Journal: Microeconomics* 2014, 6:1-36.
79. Gneezy U, Imas A: Materazzi effect and the strategic use of anger in competitive interactions. *Proceedings of the National Academy of Sciences* 2014, 111:1334-1337.
80. Tadelis S: The power of shame and the rationality of trust. *Available at SSRN 1006169* 2007.
81. Harbaugh WT, Mayr U, Burghart DR: Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 2007, 316:1622-1625.
82. Moll J, Krueger F, Zahn R, Pardini M, de Oliveira-Souza R, Grafman J: Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc Natl Acad Sci U S A* 2006, 103:15623-15628.
83. Andreoni J: Impure altruism and donations to public goods: a theory of warm-glow giving. *The Economic Journal* 1990, 100:464-477.
•84. Malmendier U, Schmidt K: You owe me. Edited by: National Bureau of Economic Research; 2012.

This paper explores how small gifts can encourage reciprocity even when it is known that the gift is intended influence behavior. The authors model positive surprises as increasing the weight that players put on the payoffs of their coplayer, which leads to positively reciprocal behavior.

85. von Scheve C, Moldt D, Fix J, von Luede R: My agents love to conform: Norms and emotion in the micro-macro link. *Computational & Mathematical Organization Theory* 2006, 12:81-100.
86. LeVeck BL, Hughes DA, Fowler JH, Hafner-Burton E, Victor DG: The role of self-interest in elite bargaining. *Proceedings of the National Academy of Sciences* 2014, 111:18536-18541.
87. Smith VL: The two faces of Adam Smith. *Southern economic journal* 1998:2-19.