CHAPTER 14

Social Learning: Emotions Aid in Optimizing Goal-Directed Social Behavior

Oriel FeldmanHall[1], **Luke J. Chang**[2]
[1]Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI, United States;
[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, United States

INTRODUCTION

Goal-directed behavior plays an enormous role in everyday human life. Goal pursuit allows humans to navigate through life in a purposeful way, facilitating the conceptualization and achievement of highly abstract and complex goals ("I want to be a doctor when I grow up"), as well as simpler, more immediate goals ("I have a hankering for a good burger, I wonder where I can find one?"). Achieving outcomes along this spectrum of goals requires a cognitive system that can translate higher level, conceptual goals into tractable, concrete actions (Gollwitzer & Moskowitz, 1996, pp. 361–399). Decades of work have been devoted to understanding the neurocognitive system that governs such goal-directed behavior. We now know much about how we represent goals and keep them in our working memory (Badre, Satpute, & Ochsner, 2012), how we update goals (Daw, Niv, & Dayan, 2005), and what happens when conflicts arise between multiple goals (Botvinick, Cohen, & Carter, 2004). However, much less is known about how goal-directed behavior unfolds in dynamic and evolving social environments. And yet, many of our most important goals—for example, being a caring parent, teacher, or citizen of the community—have enormous social and emotional qualities.

Here we propose a psychological model that captures goal-directed behavior within the social domain. Because society and social groups function better and are more effective when individuals cooperate and help others (Tyler & Blader, 2000), one long-term superordinate goal is to maintain the well-being of the group. We argue that upholding the group's welfare can be broken down into a handful of basic social needs (e.g., preventing harm to others), which serve as a core suite of goal-directed motivations. These basic goals typically act in direct opposition to the more immediate goal to self-enhance and promote one's own welfare (i.e., one's own well-being does not always bear a one-to-one correspondence with the group's well-being) (Thibaut & Kelley, 1959, 1978). The integration and subsequent resolution of these conflicting goal states is paramount to successful socialization. The affective signals (i.e., emotional prediction errors) that accompany the representations of social goals facilitate the

Goal-Directed Decision Making
ISBN 978-0-12-812098-9, https://doi.org/10.1016/B978-0-12-812098-9.00014-0

resolution of conflicts between different goals, helping to arbitrate between whether a goal is actively pursued or abandoned.

Subsequently, translating these goals into actionable outputs requires that an individual dynamically learn and update their goals through experience. Individuals constantly shift goals and strategies according to the outcomes of their previous choices and other information received from the changing environment. Since people are motivated to adapt their choices to correspond with the ever evolving social world, we posit that goal-directed behavior is flexible across contexts. In other words, social goals, and their implementation, are not stable, but rather are modulated by the context in which the goal-directed behavior arises. Finally, we conclude by proposing a social value cost function model that, depending on how each basic social need is weighted, can dictate the type of goal-directed behavior that is employed. This model facilitates a formalized testable hypothesis of the mechanisms underlying goal-directed social behavior.

Motivation and representation of social goals

Successfully navigating the social world requires that one's goals be represented in a manner consistent with promoting social well-being. Social goals are internal mental representations that relate to attaining an end state involving the welfare of others or the group (Kruglanski & Webster, 1996). Because social goals are so intimately linked with individual experiences, they can vary widely between people (Reeve, 2008). Here we argue, however, that irrespective of an individual's experience, there are a handful of goals—that manifest as needs and desires—which help to facilitate successful socialization. This set of goals are internal states that motivate behavior and goal pursuit within social environments (Ajzen & Fishbein, 1969). We suggest there are three basic goals that operate most potently within the social domain: (1) preventing harm to others, (2) social affiliation, and (3) minimizing uncertainty. This set of social goals can, at times, act in direct opposition to enhancing one's own well-being (a conflicting goal). The result is a tension between behaving in ways that facilitate successful socialization and acting in ways that augment one's own welfare.

Harm prevention

One of the most deeply held social goals is to prevent harm to others (Haidt, 2012). This desire to prevent harm has a long evolutionary lineage that can be traced back through our phylogenetic tree to ancestors common to other primate species (de Waal, 1997). Research illustrates that people are not only averse to performing harmful actions (Cushman, Gray, Gaffey, & Mendes, 2012; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Mikhail, 2000), but will go to great lengths to avoid harming another, even in the face of a superordinate goal—such as being commanded by an army officer to kill during battle (Grossman, 1996). Indeed, observing another in pain is enough to

increase one's physiological reactivity (Cushman et al., 2012), an indication of being in a highly aversive state. Individuals also report increased feelings of psychological distress in the wake of applied harm (Batson, Van Lange, Ahmad, & Lishner, 2003). This heightened aversive emotional state is generated even in the absence of actual harmful outcomes, for instance, when using a rubber knife to simulate stabbing (Cushman et al., 2012). The "avoid harm" goal manifests both as a need to refrain from impulses that may result in harm to others and as a desire to actively behave in ways that thwart harm (Bandura, 1991). Effectively, this potently experienced "avoid harm" goal helps to buoy collective social welfare and successful socialization.

Individuals who routinely exhibit behavioral patterns consistent with breaking the "avoid harm" goal (e.g., psychopaths), fail to inhibit impulses to harm, and exhibit little remorse, guilt, or empathy in the aftermath of harmful actions. Most diagnosed psychopaths do not farewell in society (Petherick, 2014), as they are three times as likely to partake in violent, criminal behavior and spend a significant portion of their adult lives behind bars (Kiehl & Hoffman, 2011). One theory posits that psychopaths do not hold the "avoid harm" goal because they exhibit generally low levels of affective physiological responsivity (Wang, Baker, Gao, Raine, & Lozano, 2012). Without sufficient levels of physiological responding, psychopaths seek out behavior that stimulates their heart rate and arousal levels, which can often take the form of violent behavior.

Affiliation

A second overarching social goal is the need to affiliate with others (Baumeister & Leary, 1995; McClelland, 1985). Examples from both inside and outside the laboratory reveal the strength by which humans feel and act on the desire to belong. Even the existence of a superficial social connection to another person (e.g., sharing a birthday) or group (e.g., finding out that you belong to an arbitrarily defined minimal group) makes the goal of needing to belong more accessible and relevant, and can ultimately bolster the motivation to affiliate (Walton, Cohen, Cwir, & Spencer, 2012).

Affiliative motivations can be so strong that an individual may conform to the behavioral patterns of others, even if the individual does not agree or approve of their own conforming behavior (Asch, 1956). In some instances, conforming to the group can be quite innocuous, such as when individuals shift their own preferences for how much they like certain musical styles (Berns & Moore, 2011; Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010) or how attractive they find another individual (Klucharev, Hytonen, Rijpkema, Smidts, & Fernandez, 2009; Zaki, Schirmer, & Mitchell, 2011). In other cases, conforming can have positive social consequences, such as enhancing one's own cooperative or charitable behavior after watching others behave prosocially (Fowler & Christakis, 2010; Nook, Ong, Morelli, Mitchell, & Zaki, 2016).

This canon of work illustrates that explicit social influence can readily induce compliant behavior (Asch, 1956; Cialdini & Goldstein, 2004). Emerging research reveals just how deep the need to affiliate goes: people are so sensitive to subtle social dynamics that they are even willing to alter their own behavior in the absence of overt social influence. For example, simply observing risk-taking behaviors in others increases one's own risk-taking profile (Suzuki, Jensen, Bossaerts, & O'Doherty, 2016). Within the moral domain, learning to implement the punishment preferences of a person who was the victim of a fairness violation enhances one's own desire to punish once they are affronted with a similar moral violation (FeldmanHall, Otto, & Phelps, 2018). Together, these findings indicate the strength by which individuals desire to affiliate and gain approval from others (Baumeister & Leary, 1995).

Converging evidence of affiliative behavior can also be observed in real-time laboratory interactions, such as when two or more individuals engage in economic games. In one-shot public goods games, where players can choose whether or not to put their own money into a common pool (which subsequently gets divided equally among the group members), contributions are typically high (Levitt & List, 2007). The fact that most people contribute, even when they understand that in order to maximize their own payout they should free ride, suggests that people are willing to forgo monetary rewards in order to signal a positive reputation to others. These findings are mirrored in one-shot prisoner's dilemmas, where 50%–60% of individuals choose to cooperate despite knowing they can potentially receive an even higher payoff if they were to defect (so long as their opponent chooses to cooperate, Barcelo & Capraro, 2015). Additional evidence from the priming literature illustrates that subconsciously activating the representation of a goal—in this case, priming participants with words related to "cooperation"—can also cause people to work together more in economic games (Bargh, Lee-Chai, Barndollar, Gollwitzer, & Trötschel, 2001).

The need to affiliate is also exemplified by how strongly individuals adhere to and enforce social norms, oftentimes at a cost to themselves. This type of goal setting is triggered by environmental cues that direct and motivate social goals to cooperate, socialize, trust, punish, and reciprocate (Charness & Rabin, 2002). If we take the case of punishment as an example, the social norm in Western cultures prescribes that a transgressing perpetrator should be punished, even if the individual who is enacting the punishment does not directly benefit herself. Many studies have demonstrated that people are willing to punish on behalf of others (i.e., third-party punishment), even when it is highly costly (Fehr & Gachter, 2002). Since it can be monetarily costly for an individual to punish a perpetrator on behalf of another victim, punishing can be likened to an altruistic act. Therefore, third-party punishment typically reflects the shared communal demands of the environment, revealing that behavior is oriented toward the collective goals of the social community and not simply the interests of the self. Indeed, without such situational cues dictating appropriate social behavior

alongside a strong desire to affiliate, individuals would likely behave in ways that would maximize their own welfare, without regard for the greater good.

Typically, the need to belong and affiliate is strongest among one's own group members: Individuals will go to great lengths to behave in ways that accord with behavioral patterns elicited from people deemed similar (e.g., conformity with ingroup but not outgroup, Cikara & Van Bavel, 2014). This need to affiliate with one's own group can be so strong that people are even willing to endure pain (in the form of electric shocks) to prevent other group members—even previously unknown individuals—from receiving electric shocks themselves (Hein, Silani, Preuschoff, Batson, & Singer, 2010).

Minimize uncertainty

Daily social life is rife with the constant need to navigate uncertain social exchanges, including deciding who to trust and confide in, whether or not to loan money to an acquaintance, or contemplating whether to participate in a potentially dangerous activity with your friends (e.g., consume alcohol or drugs). These social choices can be risky, as it is uncertain how outcomes will unfold (e.g., Can my friend keep a secret? Will my money be returned?). Most people experience uncertainty as highly aversive (Bar-Anan, Wilson, & Gilbert, 2009; Ellsberg, 1961; Hirsh, Mar, & Peterson, 2012) and therefore have a strong desire to resolve it (Kahneman, Slovic, & Tversky, 1982). These aversive feelings that accompany the experience of uncertainty become especially acute in the social domain, which acts as a potent motivator for reducing such uncertainty.

The goal to reduce social uncertainty can be met by acquiring information about other individuals. This can happen through direct experience—repeated engagements where trust or cooperative behavior can be explicitly experienced or tested (Chang, Smith, Dufwenberg, & Sanfey, 2011; Fareri, Chang, & Delgado, 2012, 2015; King-Casas et al., 2005), or by vicariously learning about the social value of another (Olsson, Nearing, & Phelps, 2007). Either route allows an individual to gain more information and update their social value estimates of other people, which leads to a stable and restricted set of expectations about the personalities and potential emotional and cognitive states of each person. In doing so, an individual can better predict what others will do, which in turn allows them to better predict their own future states. Ultimately, an individual's social success and well-being is tied to her ability to resolve the uncertainty associated with other people and social situations, with failures to do so often manifesting in clinical mood disorders (Engelmann, Meyer, Fehr, & Ruff, 2015).

Enhance self-benefit

The basic goals described above can often act in direct opposition to the desire to enhance one's own well-being and self-benefit. The ability to enhance the self—whether through money, power, or prestige—serves as an evolutionarily primal drive that aims to optimize

survival. For example, research suggests that the appeal of money can have a profoundly negative influence on people's willingness to engage in prosocial behavior. If given the opportunity to make more money through cheating and lying, individuals routinely cheat and lie (Ariely & Gino, 2011; Greene, Rand, & Nowak, 2012). Other research illustrates that when the monetary incentive is great, the number of individuals willing to break social norms, such as reciprocal trust, increases dramatically (Gneezy et al., 2011). Indeed, our own work reveals that if the monetary enhancement is sufficiently compelling (approximately $300), individuals administer electric shocks to others, willingly forgoing the "avoid harm" goal in order to enhance their own monetary self-benefit (FeldmanHall, Mobbs, Hiscox, Navrady, & Dalgleish, 2012; Feldmanhall, Dalgleish, & Mobbs, 2013).

The ease with which individuals abandon prosocial goals (e.g., affiliation, harm prevention) in the service of enhancing their own welfare, suggests that self-benefit is an equally potent and accessible goal. For example, amplifying an individual's dominance and prestige (through priming manipulations) can enhance the salience of the self-benefit goal state, which ultimately reduces the willingness to engage in prosocial acts (Guinote, 2007). Converging evidence also reveals how feeling powerful can enhance reward-seeking and disinhibited social behavior across a variety of contexts (Galinsky, Gruenfeld, & Magee, 2003; Keltner, Gruenfeld, & Anderson, 2003). In these cases, feeling powerful typically increases the rate at which an individual engages in antisocial acts, such as increased sexual aggression and harassment (Bargh, Raymond, Pryer, & Strack, 1995). In contrast, reduced power is associated with inhibited, avoidant social behavior, which traditionally aligns with group norms or the promotion of social welfare.

Translating goals to action

How do we translate social goals into actions? In order to elucidate the mechanisms governing goal-directed social behavior, we must understand the computational processes that are involved in translating superordinate goals into behavioral outputs. Quantitative disciplines such as economics, engineering, and computer science have been successful in developing normative frameworks that can describe an optimal decision policy given a set of specific goals. However, while computationally viable, these decision policies do not always capture how people *actually* behave (Camerer, 2003; Kahneman, 2003; Kahneman & Tversky, 1979). For example, standard economic theory, which assumes a rational agent is solely motivated by self-interest, is particularly poor at predicting behavior in cooperative social interactions such as bargaining (Guth, Schmittberger, & Schwarze, 1982), trust (Berg, Dickhaut, & Mccabe, 1995), and public goods games (Yamagishi, 1986). Subsequent theories that incorporate psychological motivations associated with emotions (Bell, 1982; Charness & Dufwenberg, 2006; Loewenstein, 1987, 1996; Loomes & Sugden, 1982; Mellers, Schwartz, Ho, & Ritov, 1997), concern for others' intentions (Rabin, 1993), and the collectives' outcomes

(Bolton & Ockenfels, 2000; Fehr & Gachter, 1999) have dramatically improved the ability to predict actual social behavior.

Decision policies

We review several quantitative frameworks that can be used to characterize motivated social behavior. In particular, these decision policies describe how, given an individual's social goals, selecting certain actions can maximize a stimulus' subjective value. Broadly, this process can be viewed as an optimization problem where an agent selects actions that maximize the likelihood of achieving their current social goal. We assume agents consider the expected costs and benefits associated with the outcome of a given choice and select the choice with the highest overall expected outcome for the self. There are several different types of decision policies. For example, consider a set of choice options X, where i describes a specific choice $i \in X$. Selecting the action that most aligns with the agent's goal requires applying an explicit decision policy. These policies can be deterministic, such as the greedy rule that always selects the available action a_i associated with the highest predicted expected outcome value or utility u, $a_i = argmax(u(X))$.

Alternatively, policies can select options stochastically. In these cases, choices are selected in proportion to their overall value, and the degree of stochasticity is controlled by a temperature parameter. This is traditionally modeled as a softmax function, more formally

$$a_i = \frac{e^{\frac{u_i}{\beta}}}{\sum_{j=1}^{n} e^{\frac{u_j}{\beta}}}, \tag{14.1}$$

where a_i is the probability of selecting action i, u_i is the value of action i, n is the total number of actions, and $0 < \beta < 1$ is the temperature parameter representing the stochasticity of the decision policy. Social decisions can be modeled as weighing the expected costs and benefits associated with each choice outcome and applying a decision policy to select the action with the highest overall expected utility (e.g., greedy, softmax, etc.).

Common valuation system

These various decision policies provide a principled rule for how to select an action after comparing the pros and cons of each available choice. The ability to apply a decision policy in order to carry out the desired social goal is predicated on the assumption that we can quantitatively compare the value associated with each option in the goal set. Traditionally, the pros and cons of each choice are only considered with respect to *self-interested goals* (e.g., how much money will I win or lose?). However, it remains an open question as to how we can incorporate *social goals* (e.g., harm prevention, affiliation, and social uncertainty) into a common value function, which can be compared across

choices. This requires establishing a common *value metric* (Levy & Glimcher, 2012) to integrate all of the costs and benefits associated with the available options including one's social goals (Rangel, Camerer, & Montague, 2008; Ruff & Fehr, 2014).

Economics, for example, has developed a number of tools to help establish a common value metric at the behavioral level. The Weak Axiom of Revealed Preferences (WARP) establishes the existence of a convex utility function that describes a rational agent's preferences for bundles of goods by only assuming transitivity (i.e., if A > B, and B > C, then A > C, Samuelson, 1938). Importantly, this function can also be usefully applied to social contexts, including how agents value the outcomes of others (i.e., an altruistic response, Andreoni & Miller, 2002). These are commonly referred to as social or other-regarding preferences in behavioral economics.

Expected Utility Theory is another framework for describing expectations of subjective value that adds several additional axioms to WARP in addition to transitivity (i.e., completeness, independence, and continuity) (Bernoulli, 1738; von Neumann & Morgenstern, 2007). This theory provides a powerful normative framework to describe optimal decision-making strategies when making choices under uncertainty. For example, the expected utility resulting from selecting a given choice u_i can be formally described as the sum of the value of each attribute c associated with the outcome $v_{i,c}$ scaled by the expectation of the outcome being realized $p_{i,c}$,

$$u_i = \sum_{c=1}^{n} p_{i,c} \cdot v_{i,c} \tag{14.2}$$

Though Expected Utility Theory has been very successful at providing a normative framework to understand how policies can impact economies, it has not fared as well describing how individuals make decisions. Indeed, the impressive growth and popularity of behavioral economics in the 1970s and 1980s can be attributed to an increasing realization that such normative theories of how people ought to make decisions substantially deviated from observations of how people actually behaved (Camerer, 2003; Kahneman, 2003; Kahneman & Tversky, 1979). This groundbreaking work resulted in a number of extensions to the Expected Utility Theory framework which incorporates psychological values, such as sources of value originating from social preferences (Fehr & Camerer, 2007), empathic concern for others (FeldmanHall, Dalgleish, Evans, & Mobbs, 2015), and emotional motivations (Chang & Jolly, 2017; Chang & Smith, 2015). In the following sections, we build on this framework and outline how emotional value signals arising from approaching and avoiding social goals can be integrated with self-interested value signals to impact subsequent actions.

Emotion motivates social goal-directed behavior

Emotions describe a set of phenomenological experiences that result from the intersection of our broader goals, moment-to-moment cognitive evaluations of the external

world, and our internal physiological states. Similar to the somatovisceral sensations that signal internal homeostatic goal states such as hunger, thirst, and sleep (Panksepp, 1998), emotions provide motivational signals that guide us to approach resources, avoid harm (Davidson & Irwin, 1999), and navigate the complexities of the social world (Chang & Jolly, 2017; Chang & Smith, 2015; Chang et al., 2011; FeldmanHall et al., 2013, 2015). Emotions can impact the decision-making process in a variety of ways (Chang & Sanfey, 2008; Loewenstein & Lerner, 2003). At the time of the decision, immediate emotions (e.g., gut feelings associated with risk) or incidental emotions (e.g., transient moods) can bias how we interpret information and value outcomes. Emotions can also be anticipated as an affective experience resulting from the outcome of selecting an action and incorporated directly into the value function associated with the choice. In general, we tend to value things that will make us feel good and devalue things that will make us feel bad. Ultimately, these valenced motivations can serve as signals to guide us to approach or avoid outcomes depending on our broader goals (Carver & Scheier, 1990). In the following sections, we provide examples of how emotions can impact behavior at both conscious and nonconscious levels.

Conscious emotions

Incorporating emotional motivations into utility functions through counterfactual value can dramatically improve predictions of both social (Koenigs & Tranel, 2007) and nonsocial behavior (Bell, 1982; Coricelli, Dolan, & Sirigu, 2007; Lohrenz, McCabe, Camerer, & Montague, 2007; Loomes & Sugden, 1982). Imagine buying a brand-new computer only to find out that if had you waited another week it would have been discounted an additional 15%. Though we are equally satisfied with the product, we often devalue the purchase as a consequence of regretting buying the computer a week too soon. By modeling the emotion regret, researchers have been able to capture the fact that although people are motivated by maximizing their own outcomes (e.g., the goal to enhance self-benefit), they also care about minimizing making a suboptimal decision, even if such a decision is associated with an overall positive outcome (Gilovich & Medvec, 1995; Mellers & McGraw, 2001).

An additional extension to Expected Utility Theory provided by psychological game theory is the ability to incorporate belief-dependent value into utility functions (Dufwenberg & Kirchsteiger, 2004; Geanakoplos, Pearce, & Stacchetti, 1989). This framework allows utility functions to incorporate a variety of psychological motivations, such as sensitivity to fairness (e.g., reciprocating others' good or bad intentions, Dufwenberg & Kirchsteiger, 2004; Rabin, 1993) and social emotions such as guilt from disappointing a relationship partner (Battigalli & Dufwenberg, 2007; Dufwenberg & Gneezy, 2000) and anger from believing a social norm have been violated (Battigalli, Dufwenberg, & Smith, 2015; Chang & Sanfey, 2013; Chang & Smith, 2015). The marriage of these psychological motivations with formal models has provided a useful

first-order approximation for how people integrate different sources of value and has been successfully leveraged to predict behavior in cooperative socially interactive contexts.

Nonconscious emotions

Goal-directed behavior is also known to be guided by nonconscious emotional mental processes (Aarts et al., 2005; Aarts, Custers, & Veltkamp, 2008; Custers & Aarts, 2010). In these cases, an individual's repertoire of readily available social goals is directly accompanied by a suite of implicit affective signals, typically conceptualized in terms of valence (e.g., positive or negative), that act as motivators or demotivators, depending on the context (Fazio, Sanbonmatsu, Powell, & Kardes, 1986).

For example, by activating the reward system, positive affect motivates the pursuit of approach-related social goals (Aarts et al., 2008; Custers & Aarts, 2010; Davidson, 1992). If a goal previously exists as a desired state, then it is already yoked to a positive affective signal that enhances how readily one pursues the goal—assuming the goal is primed. This has been shown to influence a number of social phenomena, including socializing (Aarts & Custers, 2007) and social equity concerns (Ferguson, 2007). Positive affect, however, can also be paired with a neutral goal (Aarts et al., 2005). In these situations, priming affectively valenced words outside of conscious awareness can activate evaluative conditioning processes (De Houwer, Thomas, & Baeyens, 2001), such that repeatedly pairing neutral goal concepts (e.g., drinking) with positive valenced phenomena (e.g., words such as "nice") increases the motivation to pursue the formally neutral goal of drinking.

In contrast, negative affect can act as a demotivator of social goal pursuit. For example, subliminally priming undergraduates with negatively valenced words (e.g., "war" or "trash") in conjunction with the goal of partying—a well-documented desired state (Sheeran et al., 2005)—made participants work less to attain the goal of partying (Aarts, Custers, & Holland, 2007). Dovetailing with this, pioneering work exploring the role of physiological arousal processes, indexed through galvanic skin responses, found that arousal levels bias how readily one continues to pursue rewarding or unrewarding choices (Bechara, Damasio, & Damasio, 2000; Ferguson & Bargh, 2004; Phelps, 2005; Winkielman & Berridge, 2004). Thus, even without conscious awareness, emotions can serve to shape choices to either pursue or abandon a social goal.

From emotions to social goals

These findings, which were popularized in canonical control theory accounts of motivated behavior (Carver, 1984; Carver & Scheier, 1981; Carver & Scheier, 1990), highlighted two critical aspects of the relationship between affect and goal pursuit in

the nonsocial domain. First, affective signals act as a basic input in determining the motivation to pursue a goal. Second, they do so by changing the rate at which an individual pursues or avoids certain goals (Cacioppo, Gardner, & Berntson, 1999; Phelps, 2005)—a theory that has proved to be robust across time and disciplines. For example, consider how we regulate our hunger. Glucose levels in our blood are constantly assessed by the hypothalamus. When glucose starts to drop below a certain homeostatic level, we begin to feel a proportional amount of hunger, which effectively prioritizes our control system to set goals to seek out food. In the pursuit of food, we employ a decision policy to find a meal that satisfies our resource constraints (e.g., location, time, and financial budget) as well as our motivational desires (that juicy burger). Once food is found and consumed, hormones begin converting the food into energy and our hunger dissipates—which then frees up our control system to prioritize other goals to pursue.

Although much less is known about how humans flexibly adapt their behavior as they navigate through social contexts that require simultaneous pursuit of multiple goals, it stands to reason that a similar relationship between emotions and goal-directed behavior exists in the social domain as well. Indeed, the notion that the affect influences social goals has been previously proposed outside the field of psychology. In sociology, for example, it has been argued that emotions can help align both our actions and identities during social interactions (Heise, 2007; Rogers, 2015; Schroder, Hoey, & Rogers, 2016). It is likely that the computational processes supporting flexible social goal pursuit are similar to other control systems that regulate human behavior (e.g., a common valuation system), including low-level homeostatic processes (Panksepp, 2004; Robinson & Berridge, 2013) and higher-level cognitive control processes (Miller & Cohen, 2001). This analogous control system allows us to incorporate and prioritize multiple, and sometimes competing, social goals. In these cases, emotions would provide both an approach and avoid signal when monitoring progress toward one of our three fundamental social goals to affiliate, minimize harm to others, or reduce our overall uncertainty—especially when these social goals come in direct conflict with the goal to enhance one's own self-benefit.

These goal-directed behaviors are also likely to be modulated by social context. For example, if resources are scarce, one's own needs are likely to be more salient—and with it, the need to enhance self-benefit. In contrast, if resources are bountiful, one is likely able to attend to the needs of others more readily. In a similar vein, if one knows they will be repeatedly encountering a certain person, goals to affiliate and minimize harm, for example, are likely to become more salient than in situations in which one encounters a person they know they will never see again. The cues provided by social environments help establish the appropriate social goal and the attendant emotional signal. Below, we outline how affective error signals might motivate us to achieve the social goals to minimize harm to others and seek out social affiliation.

Model of goal-directed social behavior

A social-affective control model

Building off accounts that illustrate affect as a key component of many goal-directed or control theories of behavior, we propose a system for selecting actions that help pursue a social goal (Fig 14.1). Critically, we believe social goals are regulated via positive and negative emotional error signals that impact our decision policies. First, the system establishes a social goal; take for instance the desire to affiliate. Values associated with different attributes of the decision space—in this case, ensuring that affiliative or conforming actions are taken—are integrated, and a decision policy (e.g., softmax) is applied to select the next action from the set of possible choices. The system then continually monitors the environment for outcomes that result from selecting this action. By evaluating how our position changes relative to the goal of social affiliation (e.g., actions that bring one closer to successfully or unsuccessfully affiliating), progress toward achieving this goal can be monitored after every action. If the action resulted in deviations from affiliating with others, this creates an error signal in the form of an emotion (e.g., negative affect, Chang & Jolly, 2017; Montague & Lohrenz, 2007). This emotional error signal is then integrated into the value function, which ultimately biases which action is selected next. The rate of change (e.g., how quickly actions are updated in accordance with the emotional error signal) reflects how quickly we are motivated toward achieving our goal to affiliate and directly corresponds to the attendant-affective responses that monitor our ongoing progress.

This framework can be used to illustrate how behavior can be optimized to maximize multiple social goals that can dynamically change and shift as our priorities also shift across various contexts. Moreover, when goals compete with one another (e.g., enhance one's own benefit or affiliate with others), the associated affective responses that monitor actions that bring one closer to either goal will shape which goal is ultimately pursued. For example, if the negative emotions that accompany failing to pursue the goal to affiliate is stronger than the negative emotional response for failing to increase the money in one's bank account, then one should pursue the goal of affiliating.

Negative affect

One way in which we can satisfy our goal to affiliate with others is to avoid acting differently from the group norm. This can be described as minimizing the distance

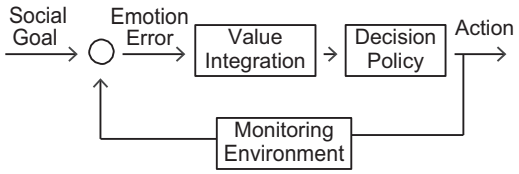

Figure 14.1 Theoretic model of how emotional error signals help us to adaptively select actions that are aligned with our social goals (e.g., preventing harm to others).

between our own behavior and shared beliefs about appropriate behavior for a specific context (i.e., social norms). Classic experiments demonstrate that we are motivated to behave in ways that consistently align with such "descriptive norms"—that is, what we believe most people would do, even when it counters our own beliefs (Asch, 1956; Cialdini & Goldstein, 2004). One instantiation of needing to affiliate with others stems from feeling intense negative affect when we are ostracized from the group or feel that we do not fit in with our peers (Leary & Richman, 2009). These negative feelings can be so powerful that the mere anticipation of being excluded, shunned, or at odds with others can generate attendant negative feelings, which in turn motivates people to act in ways that accord with the group's behavior. Within the framework of our model, the value of affiliation can be formulated as a utility function, where an agent receives utility from selecting the choice that maximizes their payoff π_i, while minimizing the deviation from the group's behavior. Here, we define a descriptive norm set by the group as the expectation of our beliefs about the likelihood of others taking certain actions i, $E[\phi_i]$ (Sanfey, Stallen, & Chang, 2014). Deviations from the group behavior generate negative affect signals.

$$u(i) = \pi_i - \alpha \cdot max(E[\phi_i] - i, 0) - \beta \cdot min(i - E[\phi_i], 0), \tag{14.3}$$

where α and β differentially scale signed deviations from the groups' normative behavior and are constrained between [0,1].

This utility function can describe a host of affiliative behaviors, including norm adherence and enforcement in the two-person bargaining task known as the ultimatum game (UG). In the UG, Player A proposes a split of an endowment to Player B. Player B then decides whether to accept the split as proposed, or reject the offer, thereby punishing Player A—in which case both players receive nothing (Guth et al., 1982). Chang and Sanfey tested a variant of this conformity model using participants' self-reported beliefs about the type of proposals they expected to encounter in the game (Chang & Sanfey, 2013). This model provided a better account of players' decisions compared to another social preference model that proposes that players prefer equitable outcomes (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999).

In a related experiment, Xiang and colleagues provide even stronger evidence for how we track social norms (Xiang, Lohrenz, & Montague, 2013). In this study, the experimenters manipulated participants' expectations in the UG by exposing different participants to three different distributions of offers (high, medium, low). In the test phase, all groups of participants were given offers from the medium distribution. Participants decided whether to accept or reject the offers and reported their affective responses by selecting from a set of emoticons. The authors combined an ideal Bayesian observer model with a social preference utility function to track how beliefs about the social norm are updated after each offer (Chang & Sanfey, 2013; Fehr & Schmidt, 1999). This provided trial-by-trial estimates of the prediction error and variance

prediction error for a given offer conditional on prior beliefs. Identical offers were rejected more frequently when participants expected offers from a high distribution compared to when participants expected offers from a low distribution, indicating that the social norm manipulation successfully changed attitudes toward fairness violations and subsequent decisions to refrain from punishing norm violators. Effectively, specific social norms can influence expectations, which in turn allow individuals to update their social goals—in this case, deciding whether to punish.

Regret

Importantly, this social-affective control system can also operate on simulated actions and does not require that the system only update after experiencing an outcome consequent to our actions (Chang & Jolly, 2017). This is a critical feature, as it allows for anticipated feelings resulting from expected outcomes to create emotional error signals (Chiu & Montague, 2008) that provide value with an associated action (Mellers & McGraw, 2001). One example of this simulated process—and one we touched on briefly before—is anticipated regret (Coricelli et al., 2007), or when an emotional error signal results from making a decision that does not align with the goal of making the best decision. If a decision turns out to have an unfavorable outcome, the deciding agent will most likely feel disappointment. However, if an agent makes a decision and learns that they could have made an even better decision regardless of outcome favorability, they will feel regret (Bell, 1982; Loomes & Sugden, 1982; Mellers et al., 1997). Thus, regret serves as an error signal that indicates deviations from the broader goal of making an optimal social choice (e.g., deciding to help someone who may subsequently return the favor, thereby eliciting better future outcomes for yourself). Importantly, regret can be anticipated at the time of decision, which can change our valuation of the choice set, and can ultimately motivate current behavior to minimize future regret (Bault, Coricelli, & Wydoodt, 2016; Coricelli et al., 2005).

Guilt

As discussed above, another important goal for agents as they navigate their social landscape is to minimize harm to others. Guilt occurs in these interpersonal contexts when one believes they have harmed or disappointed another individual (Battigalli & Dufwenberg, 2007). Guilt is considered a prosocial emotion in that agents have a tendency to take actions that repair the relationship following social transgressions (Baumeister, Stillwell, & Heatherton, 1994; Regan, Williams, & Sparling, 1972). Like regret, even the anticipation of guilt through simulating the act of committing a transgression can provide a powerful motivation for goal-directed choices to act prosocially (Massi Lindsey, 2005). This has been successfully demonstrated in the context of honoring a relationship partner's trust in the Trust Game (Berg et al., 1995). In this game, Player A invests in Player B by transferring a portion of his initial endowment

to Player B. The investment amount is multiplied by a predetermined factor (typically fourfold), and Player B then decides how much, if any, of the multiplied investment to return to Player A.

If Player A invests money in Player B, Player B generally reciprocates by sending back some portion of the money, despite there being no advantage in doing so. In fact, if Player B wanted to solely maximize her monetary payout, she should keep all the money and return nothing to Player A, as there is no fear of reprisal in one-shot Trust Games. Why then is there overwhelming evidence that Player B routinely behaves in classically "irrational" ways by sending back the money to Player A? One possibility is the anticipated guilt that Player B would feel if she kept the money, thereby failing to uphold the social contract of trust. Indeed, models that consider other-regarding preferences such as warm-glow altruism (Andreoni, 1990), intention-based reciprocity (Dufwenberg & Kirchsteiger, 2004; Rabin, 1993) and inequity-aversion (Charness & Rabin, 2002; Falk, Fehr, & Fischbacher, 2008; Fehr & Schmidt, 1999) reveal that minimizing anticipated guilt provides a powerful signal to motivate honoring a partner's trust (Battigalli & Dufwenberg, 2007; Chang & Smith, 2015; Charness & Dufwenberg, 2006; Dufwenberg & Gneezy, 2000)—accounts which dramatically outperform models derived from classical economic theory (Cox, 2004). According to this model, Player B receives positive value from the money they keep and negative value from the anticipated guilt of disappointing Player A by not returning any money. Player B has a second-order belief about the amount that Player A expects them to return, and any difference between the expectations that Player A may hold, and what Player B intends to return, can create anticipatory guilt in Player B that shapes their goal to reciprocate. Formally, Player B's utility function U_B for selecting action i can be described as

$$U_B(i) = \pi_B(i) - \theta_B \cdot \max\left(E_B^2[\pi_A] - \pi_A(i), 0\right), \tag{14.4}$$

where $\pi_B(i)$ is B's payoff for action i, $\pi_A(i)$ is A's payoff for action i, $E_B^2[\pi_A]$ is B's second-order belief about what they believe A expects his payoff to be, and θ_B is a free parameter representing Player B's sensitivity to feeling guilt. In this formalization, guilt has negative value and is represented as an emotional error signal resulting from disappointing a partner's expectations.

There have been several laboratory studies providing support for guilt-aversion stimulating prosocial behavior (Charness & Dufwenberg, 2006; Nihonsugi, Ihara, & Haruno, 2015). The amount of money that Player B returns is directly proportional to their beliefs about Player A's expectations (Dufwenberg & Gneezy, 2000) such that Player B will be even more likely to reciprocate if he/she believes that his/her partner has expectations of cooperation (Chang et al., 2011; Reuben, Sapienza, & Zingales, 2009). The fact that guilt can induce strategy changes during these tasks reveals how emotional prediction errors make individuals flexible in their choice selection—able to adaptively respond during dynamic social exchanges to achieve a common social goal.

CONCLUSIONS

How humans pursue goals has been a topic of great interest for many decades. We now know much about how goals are represented and translated into action. Much less research, however, has focused on how goal-directed behavior unfolds during social interactions. And yet, many of our most important and primary goals are qualitatively social in nature. Indeed, the success of one's social experience is directly linked with how readily one achieves their social goals. Here, we argue there are three social goals—preventing harm, affiliating with others, and resolving social uncertainty—that serve as the basic building blocks of promoting social well-being. However, these social goals often come into direct conflict with an equally potent goal—the desire to enhance our own well-being.

How humans translate these oftentimes conflicting goals into concrete actions requires learning about the world and updating the relevance of each goal according to the outcomes of previously taken actions. Here, we argue that errors in pursuing these goals can manifest in the form of specific emotional feelings such as guilt, regret, or anger, and can provide a powerful motivational signal for how people update their goals and alter their actions. These emotional error signals result from monitoring how our actions help us progress toward (or away from) a goal, acting to regulate our behavior to be consistent with our social goals akin to a control theoretic system. Accordingly, our social-affective control model offers a tractable framework for understanding how these emotional prediction errors might guide individuals to undertake or circumvent certain social behaviors during social goal pursuit. Indeed, such a formal model offers both testable hypotheses about when and how these emotional prediction errors shape social goal pursuit and provides a useful roadmap for developing future experiments that can better parse the role of emotion in guiding certain social goals.

While we primarily focused on how various negative emotional phenomena act as prediction errors to guide social goal pursuit, it is likely that there are several other positive emotional experiences, such as empathy, that also provide an error signal to promote prosocial goals (FeldmanHall et al., 2015; Lockwoord, Apps, Valton, Viding, & Roiser, 2016). Although there is little work that we are aware of that formally quantifies this process, we hypothesize that a prediction error likely motivates empathy in a similar way to those previously described above. Future work aimed at elucidating this process will deepen our understanding of the relationship between emotional experiences and social goal pursuit.

Finally, there is relatively little work that has explored how social contexts bias the relationship between emotions and goal pursuit. And yet, there are arguably many cases in which one's social goals to affiliate, minimize uncertainty or prevent harm are more salient than others. Take for example situations in which one is caring for a child versus entertaining friends. In these cases, the actions taken surrounding the care for a child will likely prioritize harm prevention, whereas the salient actions when engaging with friends

will likely correspond with the social goal to affiliate. In other words, the desire to minimize harm is not stable across all social environments but is rather modulated by the context in which the goal-directed behavior arises. Our hope is that future work can help identify and document the various contexts that either increase or decrease the selection of certain social goals.

REFERENCES

Aarts, H., Chartrand, T. L., Custers, R., Danner, U., Dik, G., Jefferis, V., & Cheng, C. M. (2005). Social stereotypes and automatic goal pursuit. *Social Cognition, 23*, 464–489.

Aarts, H., & Custers, R. (2007). In search of the nonconscious sources of goal pursuit: Accessibility and positive affective valence of the goal state. *Journal of Experimental Social Psychology, 43*, 312–318.

Aarts, H., Custers, R., & Holland, R. W. (2007). The nonconscious cessation of goal pursuit: When goals and negative affect are coactivated. *Journal of Personality and Social Psychology, 92*, 165–178.

Aarts, H., Custers, R., & Veltkamp, M. (2008). Perception in the service of goal pursuit: Motivation to attain goals enhances the perceived size of goal-instrumental objects. *Social Cognition, 26*(6), 720–736.

Ajzen, I., & Fishbein, M. (1969). Prediction of behavioral intentions in a choice situation. *Journal of Experimental Social Psychology, 5*(4), 400–416.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal of Nepal, 100*(401), 464–477.

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*, 737–753.

Ariely, D., & Gino, F. (2011). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology, 102*, 445–459.

Asch, S. E. (1956). Studies of independence and conformity .1. A minority of one against a unanimous majority. *Psychological Monographs, 70*(9), 1–70.

Badre, D., Satpute, A. B., & Ochsner, K. N. (2012). The neuroscience of goal-directed behavior. In H. Aarts, & A. J. Elliot (Eds.), *Goal-directed behavior*. New York: Taylor & Francis Group, LLC.

Bandura, A. (1991). Social cognitive theory of moral thought and action. In *Handbook of moral behavior and development, Vol. 1: Theory*. Lawrence Erlbaum Associates, Inc.

Bar-Anan, Y., Wilson, T. D., & Gilbert, D. T. (2009). The feeling of uncertainty intensifies affective reactions. *Emotion, 9*(1), 123–127.

Barcelo, H., & Capraro, V. (2015). Group size effect on cooperation in one-shot social dilemmas. *Scientific Reports, 5*.

Bargh, J. A., Lee-Chai, A., Barndollar, K., Gollwitzer, P. M., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81*(6).

Bargh, J. A., Raymond, P., Pryer, J. B., & Strack, F. (1995). Attractiveness of the underling – an automatic power→sex association and its consequences for sexual harassment and aggression. *Journal of Personality and Social Psychology, 68*(5), 768–781.

Batson, C. D., Van Lange, P. A. M., Ahmad, N., & Lishner, D. A. (2003). *Altruism and helping behavior. The sage handbook of social psychology*. Sage Publications.

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

Battigalli, P., Dufwenberg, M., & Smith, A. (2015). *Frustration and anger in games*.

Bault, N., Coricelli, G., & Wydoodt, P. (2016). Different attentional patterns for regret and disappointment: An eye-tracking study. *Journal of Behavioral Decision Making*.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*(3), 497–529.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115*(2), 243–267.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295–307.

Bell, D. E. (1982). Regret in decision-making under uncertainty. *Operations Research, 30*(5), 961–981.

Berg, J., Dickhaut, J., & Mccabe, K. (1995). Trust, reciprocity, and social-history. *Games and Economic Behavior, 10*(1), 122–142.

Bernoulli, D. (1738). *Specimen theoriae novae de mensura sortis*.

Berns, G. S., & Moore, S. E. (2011). A neural predictor of cultural popularity. *Journal of Consumer Psychology, 22*.

Bolton, G. E., & Ockenfels, A. (2000). A theory of equity, reciprocity, and competition. *American Economic Review, 90*, 166–193.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*(12), 539–546.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology, 76*(5), 839–855.

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

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology, 20*(13), 1165–1170.

Carver, C. S., & Scheier, M. F. (1981). A control-systems approach to behavioral self regulation. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 107–140). Beverly Hills CA: Sage.

Carver, C. S., & Scheier, M. F. (1984). A control-theory approach to behavior and some implications for social skills training. In P. Trower (Ed.), *Radical approaches to social skills training* (pp. 144–179). London/New York: Croom Helm/Methuen.

Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect – a control-process view. *Psychological Review, 97*(1), 19–35.

Chang, L. J., & Jolly, E. (2017). Emotions as computational signals of goal error. *Nature of Emotions*.

Chang, L. J., & Sanfey, A. G. (2008). Emotion, decision-making, and the brain. *Neuroeconomics*, 31–53. D.H.K. McCabe, Elsevier.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*(3), 277–284.

Chang, L. J., & Smith, A. (2015). Social emotions and psychological games. *Current Opinion in Behavioral Sciences*.

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560–572.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica, 74*(6), 1579–1601.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117*(3), 817–869.

Chiu, P. H., Lohrenz, T. M., & Montague, P. R. (2008). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuroscience, 11*(4), 514.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9*(3), 245–274.

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: A neuroimaging study of choice behavior. *Nature Neuroscience, 8*(9), 1255.

Coricelli, G., Dolan, R. J., & Sirigu, A. (2007). Brain, emotion and decision making: The paradigmatic example of regret. *Trends in Cognitive Sciences, 11*(6), 258–265.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior, 46*(2), 260–281.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2–7.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science, 329*(5987), 47–50.

Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition, 20*, 122–151.

Davidson, R. J., & Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences, 3*(1), 11–21.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.
De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*(6), 853–869.
Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior, 30*(2), 163–182.
Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268–298.
Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics, 75*(4), 643–669.
Engelmann, J. B., Meyer, F., Fehr, E., & Ruff, C. C. (2015). Anticipatory anxiety disrupts neural valuation during risky choice. *Journal of Neuroscience, 35*(7), 3085–3099.
Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—intentions matter. *Games and Economic Behavior, 62*(1), 287–303.
Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.
Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 35*(21), 8170–8180.
Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229–238.
Fehr, E., & Camerer, C. F. (2007). Social neuroeconornics: The neural circuitry of social preferences. *Trends in Cognitive Sciences, 11*(10), 419–427.
Fehr, E., & Gachter, S. (1999). Collective action as a social exchange. *Journal of Economic Behavior & Organization, 39*, 341–369.
Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817–868.
FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage, 105*, 347–356.
Feldmanhall, O., Dalgleish, T., & Mobbs, D. (2013). Alexithymia decreases altruism in real social decisions. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 49*(3), 899–904.
FeldmanHall, O., Mobbs, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434–441.
FeldmanHall, O., Otto, A. R., Phelps, E. A. (In Press). Learning moral values: another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology: General*.
Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: The effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology, 87*(5), 557–572.
Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences of the United States of America, 107*(12), 5334–5338.
Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*(3), 453–466.
Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior, 1*(1), 60–79.
Gilovich, T., & Medvec, V. H. (1995). The experience of regret – what, when, and why. *Psychological Review, 102*(2), 379–395.
Gollwitzer, P. M., & Moskowitz, G. B. (1996). *Goal effect on thought and behavior. Social psychology: Handbook of basic principles*. New York: Guilford Press.
Greene, J. D., Rand, D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.
Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives, 25*, 191–210.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.
Grossman, D. (1996). *On killing: The psychological cost of learning to kill in war and society*. Boston, Little: Brown.
Guinote, A. (2007). Power and goal pursuit. *Personality & Social Psychology Bulletin, 33*(8).

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental-analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367–388.
Ferguson, M. (2007). The Automatic Evaluation of Goals. *ACR North American Advances*.
Haidt, J. (2012). Chapter 7: The moral foundations of politics. In *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.
Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron, 68*(1), 149–160.
Heise, D. (2007). *Expressive order: Confirming sentiments in social actions*. New York: Springer.
Hirsh, J. B., Mar, R. A., & Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review, 119*(2), 304–320.
Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American Psychologist, 58*(9), 697.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 278.
Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review, 110*(2), 265–284.
Kiehl, K. A., & Hoffman, M. B. (2011). The criminal psychopath: History, neuroscience, treatment, and economics. *Jurimetrics, 51*, 355–397.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78–83.
Klucharev, V., Hytonen, K., Rijpkema, M., Smidts, A., & Fernandez, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61*(1), 140–151.
Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: Evidence from the ultimatum game. *Journal of Neuroscience, 27*(4), 951–956.
Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: 'Seizing' and 'Freezing'. *Psychological Review, 103*(2), 263–283.
Leary, M. R., & Richman, L. S. (2009). Reactions to discrimination, stigmatization, ostracism, and other forms of interpersonal rejection. *Psychological Review, 116*(2), 365–383.
Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001.
Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world. *Journal of Economic Perspectives, 21*, 153–174.
Lockwood, P., Apps, M., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechansims of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America, 113*(35), 9763–9768.
Loewenstein, G. (1987). Anticipation and valuation of delayed consumption. *Economic Journal, 97*(387), 666–684.
Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*(3).
Loewenstein, G. F., & Lerner, J. S. (2003). The role of affect in decision making. In R. Davidson, H. Goldsmith, & K. Scherer (Eds.), *Handbook of affective sciences*. Oxford University Press.
Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences of the United States of America, 104*(22), 9493–9498.
Loomes, G., & Sugden, R. (1982). Regret theory – an alternative theory of rational choice under uncertainty. *Economic Journal, 92*(368), 805–824.
Massi Lindsey, L. L. (2005). Anticipated guilt as behavioral motivation. *Human Communication Research, 31*(4), 453–481.
McClelland, D. C. (1985). *Human motivation*. Glenview, IL: Scott, Foresman and Company.
Mellers, B. A., & McGraw, A. P. (2001). Anticipated emotions as guides to choice. *Current Directions in Psychological Science, 10*(6), 210–214.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science, 8*(6), 423–429.
Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'a theory of justice'*. Ithaca: Cornell University.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review in Neuroscience, 24*, 167–202.
Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron, 56*(1).
Nihonsugi, T., Ihara, A., & Haruno, M. (2015). Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *Journal of Neuroscience, 35*(8).
Nook, E. C., Ong, D. C., Morelli, S. A., Mitchell, J. P., & Zaki, J. (2016). Prosocial conformity: Prosocial norms generalize across behavior and empathy. *Personality & Social Psychology Bulletin, 42*(8), 1045–1062.
Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience, 2*, 3–11.
Panksepp, J. (1998). Chapter 9: Energy is delight: The pleasures and pains of brain regulatory systems. In *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
Petherick, W. (2014). *Profiling and serial crime: Theoretical and practical issues*. Waltham, MA: Elsevier Inc.
Phelps, E. A. (2005). The interaction of emotion and cognition: Insights from studies of the human amygdala. In *The new unconscious*. New York: Guilford Press.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*(5), 1281–1302.
Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9*(7), 545–556.
Reeve, J. (2008). *Understanding motivation and emotion*. John Wiley & Sons, Inc.
Regan, D. T., Williams, M., & Sparling, S. (1972). Voluntary expiation of guilt: A field experiment. *Journal of Personality and Social Psychology, 24*(1), 42.
Reuben, E., Sapienza, P., & Zingales, L. (2009). Is mistrust self-fulfilling. *Economics Letters, 104*(2), 89–91.
Robinson, M. J., & Berridge, K. C. (2013). Instant transformation of learned repulsion into motivational "wanting". *Current Biology, 23*(4), 282–289.
Rogers, K. B. (2015). Expectation states, social influence, and affect control: Opinion and sentiment change through social interaction. In *, Vol. 32. Advances in group processes* (pp. 65–98). Emerald Group Publishing Limited.
Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549–562.
Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica, 5*(19), 353–354.
Sanfey, A. G., Stallen, M., & Chang, L. J. (2014). Norms and expectations in social decision-making. *Trends in Cognitive Sciences, 18*(4), 172–174.
Schroder, T., Hoey, J., & Rogers, K. B. (2016). Modeling dynamic identities and uncertainty in social interactions. *American Sociological Review, 81*(4), 828–855.
Sheeran, P., Aarts, H., Custers, R., Rivis, A., Webb, T. L., & Cooke, R. (2005). The goal-dependent automaticity of drinking habits. *The British Journal of Social Psychology, 44*(Pt 1), 47–63.
Suzuki, S., Jensen, E. L., Bossaerts, P., & O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Science of the United States of America, 113*(14), 3755–3760.
Thibaut, J. W., & Kelley, H. H. (1959). *The social psychology of groups*. New York: John Wiley & Sons, Inc.
Thibaut, J. W., & Kelley, H. H. (1978). *Interpersonal relations: A theory of interdependence*. London: John Wiley & Sons, Inc.
Tyler, T. R., & Blader, S. L. (2000). *Cooperation in groups: Procedural justice, social identity and behavioral engagement*. New York: Taylor & Francis Group, LLC.

von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton University Press.

de Waal, F. (1997). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

Walton, G. M., Cohen, G. L., Cwir, D., & Spencer, S. J. (2012). Mere belonging: The power of social connections. *Journal of Personality and Social Psychology, 102*(3), 513–532.

Wang, P., Baker, L. A., Gao, Y., Raine, A., & Lozano, D. I. (2012). Psychopathic traits and physiological responses to aversive stimuli in children aged 9-11 years. *Journal of Abnormal Child Psychology, 40*(5), 759–769.

Winkielman, P., & Berridge, K. C. (2004). Unconscious emotion. *Current Directions in Psychological Science, 13*, 120–123.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(3).

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science, 22*(7), 894–900.