83 Neural Mechanisms of Social Learning

DOMINIC S. FARERI, LUKE J. CHANG, AND MAURICIO DELGADO

ABSTRACT Our well-being is contingent upon our ability to navigate challenges and make decisions within a dynamic social environment. Social learning provides unique opportunities to meet such challenges by helping us to reduce uncertainty, update social expectations, and ultimately maximize social gains by developing close relationships. This chapter will review the mechanisms of social learning, focusing on how we can learn from and about others, how we can learn about others' mental states, and how we come to represent social relationships and social distance.

Our days are often spent navigating a complex and dynamic social environment in pursuit of various goals. For example, conducting simple transactions (e.g., buying a meal) often leads to interactions with complete strangers. We typically interact with others on a daily basis who comprise multiple interleaved social networks (e.g., family, friends, professional colleagues). Even when we are ostensibly alone, we can still be immersed in a social world when consuming media through a book, television, or the Internet. Given the preponderance of our lives spent embedded in a social context, a key question is understanding how and what types of information we learn from the social environment.

Humans have strong motivations to approach resources, while avoiding harm for self and others, and reduce uncertainty about the world (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; FeldmanHall & Chang, 2018). We are also intensely driven to form close relationships with others (Baumeister & Leary, 1995). These two overarching goals motivate much of social learning. We can accelerate reducing our uncertainty about the world by learning vicariously *from* others' experiences from both observation and direct communication. Similarly, we can also reduce our uncertainty about others by learning *about* their beliefs, motivations, preferences, and overall character—for example, how does a certain person think about the world? What types of experiences have shaped their beliefs and perspective? What type of moral character do they have and would they be a good colleague? The reduction of social uncertainty can facilitate subsequent social interactions and the development of close relationships. This chapter will review several aspects of social learning, such as how we learn: from and about others, what other people are thinking, and how people are connected to each other. Much of this work is based on basic learning concepts (e.g., Pavlovian, instrumental, goal-directed, habitual), such as forming associations between stimuli and updating beliefs based on feedback, and suggests reliance on neural circuits comprising the amygdala, dorsal and ventral striatum (DS, VS), anterior cingulate cortex (ACC), and ventromedial prefrontal cortex (vmPFC; Delgado, 2007; Haber & Knutson, 2010; O'Doherty, 2004; Phelps & LeDoux, 2005; Yin & Knowlton, 2006). However, rather than simple sensory or affective signals, this information is often gleaned through the lens of social cognition. Thus, much of the literature reviewed involves interactions between neural systems supporting learning, affect, and social reasoning.

Learning from Others

We are motivated at once to both maximize our self-interest and minimize our uncertainty about the world. This requires us to frequently switch between exploiting what we know and exploring the unknown (Cohen, McClure, & Yu, 2007). Social learning offers the advantage of minimizing our uncertainty about the world based on others' experiences without incurring our own costs from exploring. This type of fictive learning (Lohrenz, McCabe, Camerer, & Montague, 2007) could be based on simply observing the outcomes of others' actions (i.e., observational learning). Alternatively, it can be learned from directly communicating these experiences, such as being explicitly told which is the best option.

Observational learning Observing the outcomes of others while minimizing our own costs is vital for survival from the earliest stages of life. This extension of Pavlovian learning can provide key insight into the nature of threats in the environment and how to avoid them, thereby ensuring survival (reviewed in Olsson & Phelps, 2007). The observational learning of stimuli

paired with aversive outcomes results in equivalent learning as direct experience. Observationally learned cues are associated with increased physiological arousal and increased activation of the amygdala, anterior cingulate cortex (ACC), and insula (Olsson, Nearing, & Phelps, 2007). Rodent work has demonstrated that neurons projecting from the ACC, the basolateral nucleus of the amygdala (BLA), preferentially fire to cues learned via observing a conspecific undergo fear conditioning, while BLA neurons demonstrate reduced responding to such cues when ACC projections are inhibited (Allsop et al., 2018). Single-cell recordings in epilepsy patients also implicate rostral ACC neurons in the encoding of computational signals of observation, in contrast to amygdala and medial prefrontal cortex (mPFC) neurons, which show stronger involvement during firsthand experience of outcomes (Boorman, Fried, & Hill, 2016). Importantly, the extinction of a learned fear association can transmit vicariously across individuals (Golkar, Selbing, Flygare, Ohman, & Olsson, 2013), suggesting that this method of gleaning information from others aids in reducing uncertainty and avoiding harm.

Observational learning can also help us maximize gain and approach resources. For example, observing a person perform a given task can serve as an anchor (i.e., prior) that we can use to maximize our own performance based on subsequent experience. Similarly, we can make predictions about whether success will come to others and adjust our expectations after observing their outcomes. Such observational prediction error signals (i.e., expected observed outcomes) have been captured in the vmPFC, VS (Burke, Tobler, Baddeley, & Schultz, 2010), and DS (Cooper, Dunne, Furey, & O'Doherty, 2012), regions implicated in functional magnetic resonance imaging (fMRI) studies of associative and instrumental learning (Garrison, Erdeniz, & Done, 2013), as well as the intraparietal sulcus and dorsomedial prefrontal cortex (dmPFC; Dunne, D'Souza, & O'Doherty, 2016). Action prediction errors (i.e., of what others will *do*) are more associated with lateral PFC (Burke et al., 2010). Taken together, observational learning is a powerful social mechanism—through which we learn about the environment while reducing exposure to possible harm—that relies heavily on neural circuits supporting learning from direct experiences.

Social nudges Efforts to reduce uncertainty in the social world are often complicated by considerations of risk. In such situations we may look to others as a guide for whether to be risky or more prudent. Hearing from a friend or colleague who just invested in a stable rather than a more volatile stock may sway or nudge our own investments, with positive or negative consequences. Indeed, participants become risk averse when others are risk averse and become more risk seeking when others are risk seeking (Chung, Christopoulos, King-Casas, Ball, & Chiu, 2015), suggesting a *utility* placed on others' behavior that tracks with changes in vmPFC activity. This pattern of "contagion" is driven by a change in one's own risk attitudes (Suzuki, Jensen, Bossaerts, & O'Doherty, 2016). Relatedly, the vmPFC also appears to track others' *confidence* about their choice, which can influence our own decisions to pursue risk and uncertainty (Campbell-Meiklejohn, Simonsen, Frith, & Daw, 2017). These findings suggest that the overall value of these social and nonsocial signals appears to be integrated in the vmPFC and guides learning in uncertain environments (Behrens, Hunt, Woolrich, & Rushworth, 2008). Social nudges can also arise from evaluative feedback from peers, which is particularly important to consider given the dramatic rise in engagement with social media (Rodman, Powers, & Somerville, 2017). For example, even the mere presence of a peer can have an impact on reward-related neural activation (Fareri, Niznikiewicz, Lee, & Delgado, 2012), influence decisions to take risks (Chein, Albert, O'Brien, Uckert, & Steinberg, 2011), and lead to prosocial decision-making (Izuma, Saito, & Sadato, 2010), in possible anticipation of social approval. In sum, taking cues from others can significantly influence day-to-day decisions, particularly with respect to reducing uncertainty and validating our own choices.

Instructed learning A more explicit way of reducing uncertainty comes through directly receiving rules about environmental contingencies from another person. Learning via instruction is a more top-down and rapid process that can impact the goals of reducing uncertainty and maximizing one's best interest. For example, being provided (incorrect) instructed information about which of two stimuli will most likely lead to a reward will bias choice toward ostensibly more rewarding options, which hold even in the face of inconsistent feedback (i.e., punishment). Thus, explicit instruction may inhibit the appropriate updating of one's expectations (Doll, Jacobs, Sanfey, & Frank, 2009), consistent with prefrontal regulation of instrumental striatal learning processes (Li, Delgado, & Phelps, 2011). Instructions can also impact our ability to learn to avoid harm via corticostriatal circuitry during reversal learning (Atlas, Doll, Li, Daw, & Phelps, 2016). Interestingly, instructions from others concerning the *reliability* of upcoming feedback may moderate these biased processes (Schiffer, Siletti, Waszak, & Yeung, 2017).

Learning about Others

In addition to reducing uncertainty about the world, we are also motivated to build relationships and forge connections with others. This requires building a model of a person that can predict their behavior across a range of contexts (e.g., how good or trustworthy is this person?). We can then update this model based on simple information about a person's social relations and group membership through direct interactions or vicariously through another person's experience. More sophisticated models might incorporate information about an agent's personality, preferences, or how the agent thinks about the world—that is, the agent's beliefs, desires, and intentions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017).

Trait learning and impression updating We often form simple models of others by trying to infer their traits. Upon meeting someone novel, we might make implicit judgments about their level of trustworthiness or approachability based on facial characteristics (Todorov, Baron, & Oosterhof, 2008), assumed knowledge of their affiliations with a particular social group (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011), or their beliefs about the world (i.e., stereotypes; Freeman & Johnson, 2016). These snap judgments contribute to the initial models we construct about others based on social approach and avoidance motives (Willis & Todorov, 2006). Forming first impressions implicates the amygdala (Engell, Haxby, & Todorov, 2007) and posterior cingulate cortex (PCC) in representing valenced social information, as well as the dmPFC in representing more general information about a person (Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009).

Navigating our social landscapes requires constantly updating our initial models of others. We can do this readily when we acquire new information about a person that is perceived to occur with high statistical frequency in the social environment (i.e., more people tend to act trustworthy than not; Mende-Siedlecki, Baron, & Todorov, 2013). The dmPFC, PCC, and superior temporal sulcus (STS), all regions supporting social cognition (Stanley & Adolphs, 2013), are especially important for tracking inconsistencies in diagnostic social information about a target (Mende-Siedlecki, Cai, & Todorov, 2012). Further, positive changes in impressions (based on information about competence) may be mediated by increasing activation in lateral PFC, while negative changes in impressions of competence tend to recruit activation in mPFC, the striatum, and the STS (Bhanji & Beer, 2013).

Social interactions and reputation First impressions serve as a baseline expectation of other individuals that inform the likelihood of future successful interactions with them. Violations of social expectations (e.g., thinking we will be liked, only to find out we are not) tend to recruit regions involved in processing cognitive conflict and error monitoring, such as the dorsal ACC, whereas the ventral ACC discriminates between the valence of social outcomes agnostic to initial expectations (Cooper, Dunne, Furey, & O'Doherty, 2014; Somerville, Heatherton, & Kelley, 2006). The encoding of such signals in the ACC, VS, and mPFC provide neural mechanisms through which we can learn about social targets likely to provide opportunities for social inclusion and affiliation during repeated interactions (Jones et al., 2011).

Repeated interactions with a partner enable learning about *reputation*, which facilitates the development of relationships (Fareri & Delgado, 2014b). Trust underscores learning about one's reputation and can be operationalized as the expectation that someone will reciprocate generosity in situations involving mutual, interdependent risk (Simpson, 2007). Reciprocity serves as a valued social commodity that is consistently represented in corticostriatal reward systems (Bellucci, Chernyak, Goodyear, Eickhoff, & Krueger, 2016; Phan, Sripada, Angstadt, & McCabe, 2010). Experienced reciprocity during repeated interactions with a partner significantly predicts whether we should continue to collaborate with someone, as peak blood oxygen level-dependent (BOLD) activation in the caudate nucleus exhibits a temporal shift from the time at which a partner's choice to reciprocate is revealed to an anticipatory peak prior to the revelation of a partner's response (King-Casas et al., 2005). This pattern of striatal activation is consistent with temporal difference learning models that have been reported in midbrain dopaminergic neurons of nonhuman primates (Hollerman & Schultz, 1998), suggesting a social reward prediction error that can aid in updating social expectations/reputation. Expectations of reciprocity are susceptible to outside influence (i.e., prior instructed information about a partner's moral character): people tend to trust those of positive moral character over those of negative moral character, even when faced with information inconsistent with said priors (Delgado, Frank, & Phelps, 2005). This phenomenon may be driven by the interference of instructed social priors with striatal learning mechanisms to appropriately update social expectations.

Computational mechanisms of impression updating Updating social impressions is thus a dynamic process

requiring a comparison of initial expectations/impressions and current experiences (Chang, Doll, van 't Wout, Frank, & Sanfey, 2010), and recent years have seen a steady increase in the incorporation of computational approaches to learning about others. Reinforcement-learning (RL) approaches (Dayan & Daw, 2008; Sutton & Barto, 1998), for example, offer opportunities to apply additional precision to social neuroscientific questions via the mathematical formalization of specific hypotheses regarding social behavior (Cheong, Jolly, Sul, & Chang, 2017). The recent application of RL models to learning about others has delineated neurocomputational mechanisms supporting trait versus reward learning. When faced with the task of choosing between social targets that could share some portion of an endowment, participants appear to use information about outcomes (i.e., amount shared) and generosity (i.e., what was the total amount *available* to be shared by someone) to inform choice and learning (Hackel, Doll, & Amodio, 2015). This study also reported overlapping activation in the VS for learning signals associated with both reward and generosity, consistent with extant research (Garrison, Erdeniz, & Done, 2013), but generosity also recruited a network of putative social regions (PCC, precuneus and right temporoparietal junction [rTPJ]). A related study found that learning about an individual's traits could be described using the same Bayesian model as learning about monetary reward, but the neurocomputational signals supporting social learning are encoded almost exclusively in putative social regions (i.e., precuneus; Stanley, 2016).

RL approaches have also been applied to studies examining trust and reputation learning. Models assuming that trust is a dynamic process posit that initial impressions shape the manner in which new information is incorporated into belief updating about another individual (Chang et al., 2010). Indeed, if initial impressions are strong enough, they can influence how much we subsequently value and use reciprocity/defection to learn about a partner. When priors acquired through direct social experience exist about another person, individuals show higher learning rates for outcomes that are *consistent* with initial impressions than for outcomes that are inconsistent, demonstrating that prior expectations computationally influence impression updating (Fareri, Chang, & Delgado, 2012). Strong instructional priors also modulate the neurocomputational mechanisms of social learning. During violations of trust, connectivity between the striatum and ventrolateral prefrontal regions is enhanced when priors are present, suggesting inhibitory functional interactions that prevent successful impression updating (Fouragnan et al., 2013).

Learning about mental representations Inherent in our ability to use social outcomes to build a model of another's reputation is the idea that we also need to be able to understand what types of goals motivate their behavior (Baker et al., 2017). Being able to represent something about others' mental states and affective experiences (Spunt & Adolphs, 2017)—cornerstones of social cognition—is key to social learning across development, with the dmPFC supporting such computations (Sul, Guroglu, Crone, & Chang, 2017). Multivariate analyses reveal that neural networks that support mentalizing represent information about others' mental states along three key dimensions—rationality (dmPFC, anterior temporal lobe), social impact or relevance (TPJ, precuneus, rostral ACC, dorsal ACC [dACC]), and valence (TPJ, dlPFC, inferior frontal gyrus/insula; Tamir, Thornton, Contreras, & Mitchell, 2016). These dimensions of mental state representation are critically involved in the ability to predict the manner in which individuals will transition between similar/different emotional states, something that overall we tend to be able to predict with high degrees of accuracy (Thornton & Tamir, 2017). In addition, modeling others' mental states requires reasoning about how others will interpret and respond to our actions. Complex computational strategies instantiated in the mPFC and STS (and supported by interactions with the VS) indeed track both another's (e.g., teacher) actions on a trial-by-trial basis and estimations of how one's own behavior will influence the future actions of another (Hampton, Bossaerts, & O'Doherty, 2008). Further, learning about others' preferences for risky behavior (Suzuki et al., 2016) to inform our own choices relies on Bayesian mechanisms and mentalizing circuitry (e.g., dmPFC, dlPFC, inferior parietal lobule [iPL]), such that we use our own baseline preferences as a starting point from which to update beliefs about others.

Learning about social space Social interactions typically occur within rich environments with more than one person. Thus, we can derive important information about people by learning about their place within social space. Indeed, humans develop and immerse themselves in widely interconnected social networks comprised of close others, varying degrees of friends of friends, and other acquaintances. As such, this type of social learning provides information indirectly about traits and the value of others through understanding how people relate to each other within a network of individuals. For example, networks of individuals characterized by empathy tend to be those that involve closer, trusting relationships between individuals (Morelli, Ong, Makati, Jackson, & Zaki, 2017). Interestingly, social

network complexity maps on to ventrolateral and medial amygdala functional connectivity (Bickart, Hollenbeck, Barrett, & Dickerson, 2012), and other findings implicate mPFC in distinguishing representations of self and others as a function of similarity and closeness (Krienen, Tu, & Buckner, 2010; Mitchell, Macrae, & Banaji, 2006).

Other work indicates that both reward-related (VS) and social regions (mPFC) differentially integrate information about relationship closeness into value representations of in-network versus out-of-network social experiences (Fareri et al., 2012; Fareri & Delgado, 2014a). For example, collaborative interactions with close others are associated with computational signals of social reward value, represented in the VS and mPFC when experiencing reciprocity, that are contingent upon interpersonal aspects of a close relationship (Fareri, Chang, & Delgado, 2015). Relatedly, people are willing to forgo self-interest (i.e., higher monetary gain) in favor of more equitable splits with another person as a function of social closeness, a pattern that scales with activation in value-related (vmPFC) and social (rTPJ) brain regions (Strombach et al., 2015). Conversely, decisions to trust out-of-network members requires connectivity between regions implicated in cognitive control (i.e., dACC, lateral PFC) and the striatum, presumably to inhibit prepotent responses to distrust such individuals (Hughes, Ambady, & Zaki, 2016).

More recently, there has been growing interest in exploring how we learn the structure of social relationships. Judging social distance within a social network appears to recruit the same regions involved in judging spatial and temporal distance (Parkinson, Liu, & Wheatley, 2014), whereas judging the popularity of various members of a social network appears to recruit activation in reward circuitry (vmPFC, amygdala, VS) and social cognition networks (dmPFC, precuneus, left TPJ) (Zerubavel, Bearman, Weber, & Ochsner, 2015). Patterns within social cognition networks when viewing faces can also predict which members have the highest social value within a social network (i.e., sources of friendship, empathy, and support) (Morelli, Leong, Carlson, Kullar, & Zaki, 2018). Finally, there is intriguing recent evidence of neural homophily that suggests we may have more similar patterns of brain activity to our friends when viewing videos than to more distant others (i.e., friends of friends) (Parkinson, Kleinbaum, and Wheatley, 2018). Taken together, these findings suggest that shared preferences and interpretations of the world may help explain why we become closer to certain individuals than others.

Future Directions and Conclusions

Social learning serves to reduce uncertainty in the environment, maximize gains and avoid harm, and forge close relationships with others. The neural systems across many different types of social learning covered here rely heavily on interactions between corticostriatal circuitry and the cortical regions supporting social processing (Figure 83.1).

We note that the topics covered here are not exhaustive. For instance, social learning can occur via other means, such as through the adherence to and enforcement of social norms (Chang & Sanfey, 2013; Montague & Lohrenz, 2007; Xiang, Lohrenz, & Montague, 2013; Zaki, Schirmer, & Mitchell, 2011; Zhong, Chark, Hsu, & Chew, 2016) or the desire to avoid feelings of guilt for committing social transgressions (Chang, Smith, Dufwenberg, & Sanfey, 2011; Nihonsugi, Ihara, & Haruno, 2015).

With respect to future directions, one exciting path concerns more concrete models of observational learning—that is, does this type of learning occur simply via the simple imitation of an agent or rather through using our observations of others to generate a model about environmental states (i.e., inverse reinforcement learning; Collette, Pauli, Bossaerts, & O'Doherty, 2017)? Better characterizing observational learning mechanisms can foster a deeper understanding of theory of mind processes and how they may break down in clinical samples (i.e., autism). Another interesting direction involves harnessing machine-learning algorithms to facilitate the prediction of psychological states (i.e., negative affect) based on decoding patterns of brain activation (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015). Translating these types of predictive techniques to questions of social appraisals (i.e., reputation, bias) and social decisions (i.e., trust) has implications for understanding breakdowns in representations of others with interpersonal difficulties. Finally, developing stable, long-lasting relationships depends heavily upon the processes reviewed in this chapter. Learning about and from others facilitates the development and maintenance of close, trusting relationships, which supports our overall well-being (Uchino, 2009). Future work can take a more comprehensive approach to characterizing the dynamics of relationships and shared experiences across groups of individuals as they relate to processing, learning, and remembering social information in more naturalistic contexts (Chen et al., 2016) and how this subsequently influences mental health.

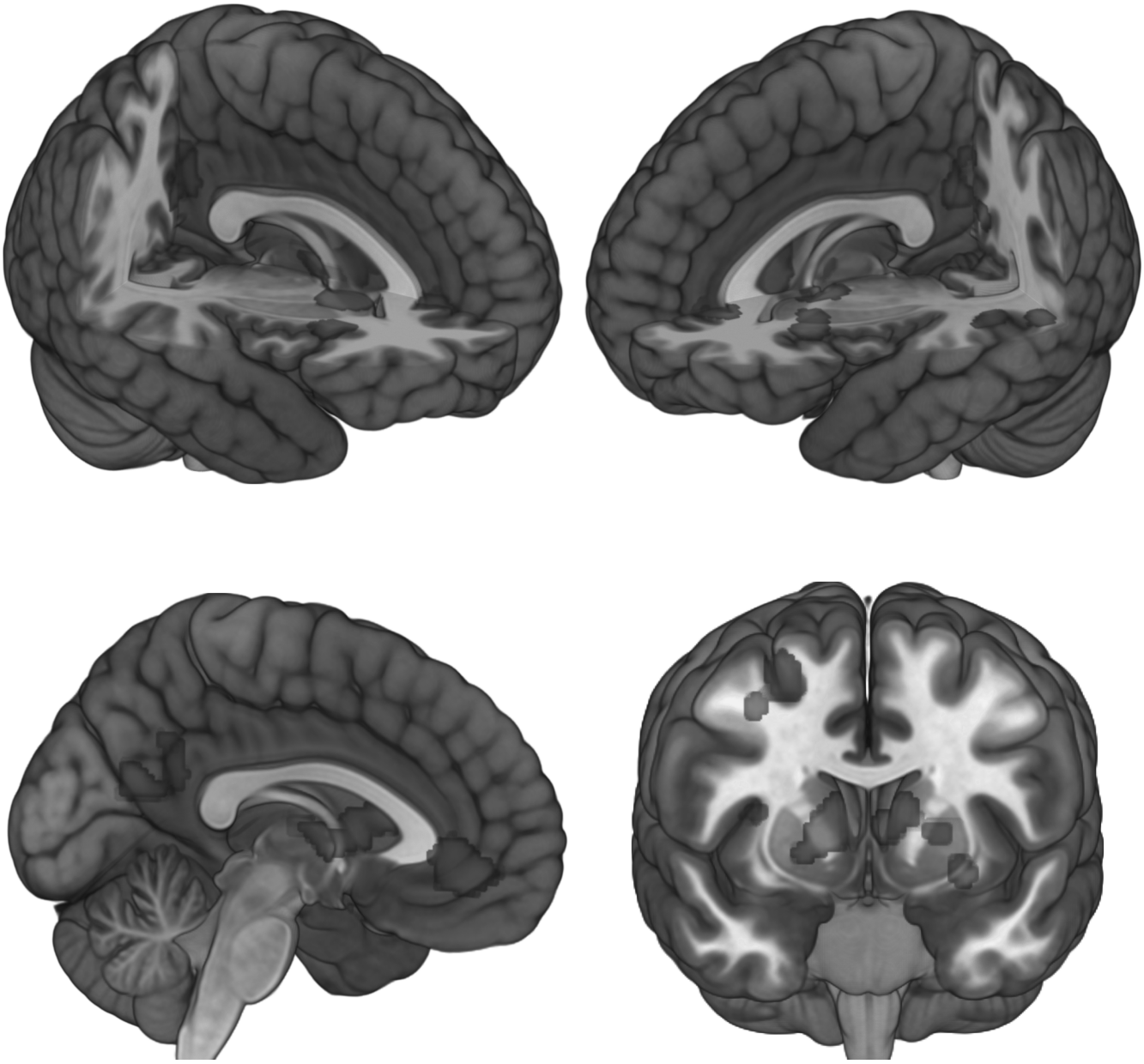

FIGURE 83.1 Activation-likelihood meta-analyses using GingerALE (Eickhoff et al., 2009) were conducted to generate illustrative maps of neural circuitry supporting "learning from" (*green*) and "learning about" (*red*) others. Maps were set to an initial height threshold of $p < .005$ and corrected at the cluster level to $p < .05$. Studies included in these meta-analyses are marked with an * (*learning from*) and a † (*learning about*) in the reference section. (See color plate 96.)

REFERENCES

Achterberg, M., van Duijvenvoorde, A. C. K., Bakermans-Kranenburg, M. J., & Crone, E. A. (2016).* Control your anger! The neural basis of aggression regulation in response to negative social feedback. *Social Cognitive and Affective Neuroscience, 11*(5), 712–720. http://doi.org/10.1093/scan/nsv154

Allsop, S. A., Wichmann, R., Mills, F., Burgos-Robles, A., Chang, C.-J., Felix-Ortiz, A. C., et al. (2018). Corticoamygdala transfer of socially derived information gates observational learning. *Cell, 173*(6), 1–33. http://doi.org/10.1016/j.cell.2018.04.004

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016).* Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *eLife, 5*, e15192. http://doi.org/10.7554/eLife.15192

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behavior, 1*, 1–10. http://doi.org/10.1038/s41562-017-0064

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

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008).* Associative learning of social value. *Nature, 456*(7219), 245–249. http://doi.org/10.1038/nature07538

Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., & Krueger, F. (2016). Neural signatures of trust in reciprocity: A coordinate-based meta-analysis. *Human Brain Mapping, 38*(3), 1233–1248. http://doi.org/10.1002/hbm.23451

Bhanji, J. P., & Beer, J. S. (2013).† Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. *Journal of Neuroscience, 33*(22), 9337–9344. http://doi.org/10.1523/jneurosci.5634-12.2013

Bickart, K. C., Hollenbeck, M. C., Barrett, L. F., & Dickerson, B. C. (2012). Intrinsic amygdala-cortical functional connectivity predicts social network size in humans. *Journal of Neuroscience, 32*(42), 14729–14741. http://doi.org/10.1523/jneurosci.1599-12.2012

Boorman, E. D., Fried, I., & Hill, M. R. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications, 7*, 1–12. http://doi.org/10.1038/ncomms12722

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010).* Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32). http://doi.org/10.1073/pnas.1003111107

Cacioppo, S., & Cacioppo, J. T. (2019). Mechanisms of loneliness. In M. S. Gazzaniga, G. R. Mangun, & D. Poeppel (Eds.), *The cognitive neurosciences* (6th ed.). Cambridge, MA: MIT Press.

Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D., & Daw, N. D. (2017).* Independent neural computation of value from other people's confidence. *Journal of Neuroscience, 37*(3), 673–684. http://doi.org/10.1523/jneurosci.4490-15.2016

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105. http://doi.org/10.1016/j.cogpsych.2010.03.001

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biology, 13*(6), e1002180–28. http://doi.org/10.1371/journal.pbio.1002180

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. http://doi.org/10.1093/scan/nsr094

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

Chein, J., Albert, D., O'Brien, L., Uckert, K., & Steinberg, L. (2011). Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. *Developmental Science, 14*(2), F1–10. http://doi.org/10.1111/j.1467-7687.2010.01035.x

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2016). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience, 20*(1), 115–125. http://doi.org/10.1038/nn.4450

Cheong, J. H., Jolly, E., Sul, S., & Chang, L. J. (2017). Computational models in social neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 1–16). Hoboken, NJ: John Wiley & Sons.

Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., & Chiu, P. H. (2015).* Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience, 18*(6), 912–916. http://doi.org/10.1038/nn.4022

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 933–942. http://doi.org/10.1098/rstb.2007.2098

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife, 6*. http://doi.org/10.7554/eLife.29718

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012).* Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience, 24*(1), 106–118. http://doi.org/10.1162/jocn_a_00114

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2014). The role of the posterior temporal and medial prefrontal cortices in mediating learning from romantic interest and rejection. *Cerebral Cortex, 24*(9), 2502–2511. http://doi.org/10.1093/cercor/bht102

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America, 111*(48), 17320–17325. http://doi.org/10.1073/pnas.1408988111

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

Delgado, M. R. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences, 1104*, 70–88. http://doi.org/10.1196/annals.1390.002

Delgado, M. R., Frank, R., & Phelps, E. A. (2005).† Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*, 1611–1618. http://doi.org/doi:10.1038/nn1575

Dewall, C. N., Deckman, T., Gailliot, M. T., & Bushman, B. J. (2011). Sweetened blood cools hot tempers: Physiological self-control and aggression. *Aggressive Behavior, 37*(1), 73–80. http://doi.org/10.1002/ab.20366

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research, 1299*, 1–21. http://doi.org/10.1016/j.brainres.2009.07.007

Dunne, S., D'Souza, A., & O'Doherty, J. P. (2016).* The involvement of model-based but not model-free learning signals during observational reward learning in the absence of choice. *Journal of Neurophysiology, 115*(6), 3195–3203. http://doi.org/10.1152/jn.00046.2016

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping, 30*(9), 2907–2926. http://doi.org/10.1002/hbm.20718

Engell, A. D., Haxby, J. V., & Todorov, A. (2007).† Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508–1519. http://doi.org/10.1162/jocn.2007.19.9.1508

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*, 148–217. http://doi.org/10.3389/fnins.2012.00148

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015).† Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience, 35*(21), 8170–8180. http://doi.org/10.1523/jneurosci.4775-14.2015

Fareri, D. S., & Delgado, M. R. (2014a). Differential reward responses during competition against in- and out-of-network others. *Social Cognitive and Affective Neuroscience, 9*(4), 412–420. http://doi.org/10.1093/scan/nst006

Fareri, D. S., & Delgado, M. R. (2014b). The importance of social rewards and social networks in the human brain. *Neuroscientist, 20*(4). http://doi.org/10.1177/1073858414521869

Fareri, D. S., Niznikiewicz, M. A., Lee, V. K., & Delgado, M. R. (2012). Social network modulation of reward-related signals. *Journal of Neuroscience, 32*(26), 9045–9052. http://doi.org/10.1523/jneurosci.0610-12.2012

FeldmanHall, O., & Chang, L. J. (2018). Social learning: Emotions aid in optimizing goal-directed social behavior. In A. M. Bornstein & A. Shenhav (Eds.), *Understanding goal-directed decision-making computations and circuits.* Amsterdam: Elsevier.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013).† Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience, 33*(8), 3602–3611. http://doi.org/10.1523/jneurosci.3086-12.2013

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences, 20*(5), 362–374. http://doi.org/10.1016/j.tics.2016.03.003

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews, 37*(7), 1297–1310. http://doi.org/10.1016/j.neubiorev.2013.03.023

Golkar, A., Selbing, I., Flygare, O., Ohman, A., & Olsson, A. (2013). Other people as means to a safe end. *Psychological Science, 24*(11), 2182–2190. http://doi.org/10.1177/0956797613489890

Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology, 35*(1), 4–26. http://doi.org/10.1038/npp.2009.129

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015).† Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233–1235. http://doi.org/10.1038/nn.4080

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008).† Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746. http://doi.org/10.1073/pnas.0711099105

Hollerman, J., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4), 304–309.

Hornstein, E. A., Inagaki, T. K., & Eisenberger, N. I. (2019). More than just friends: An exploration of the neurobiological mechanisms underlying the link between social support and health. In M. S. Gazzaniga, G. R. Mangun, & D. Poeppel (Eds.), *The cognitive neurosciences* (6th ed.). Cambridge, MA: MIT Press.

Hughes, B. L., Ambady, N., & Zaki, J. (2016).† Trusting outgroup, but not ingroup members, requires control: Neural and behavioral evidence. *Social Cognitive and Affective Neuroscience, 12*(3), 372–381. http://doi.org/10.1093/scan/nsw139

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience, 22*(4), 621–631. http://doi.org/10.1162/jocn.2009.21228

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., et al. (2011).† Behavioral and neural properties of social reinforcement learning. *Journal of Neuroscience, 31*(37), 13039–13045. http://doi.org/10.1523/jneurosci.2972-11.2011

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341–350. http://doi.org/10.1037/0003-066X.39.4.341

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

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., et al. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion, 17*(8), 1144–1155. http://doi.org/10.1037/emo0000296

Kriegeskorte, N. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 1–28. http://doi.org/10.3389/neuro.06.004.2008

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*(41), 13906–13915. http://doi.org/10.1523/jneurosci.2180-10.2010

Li, J., Delgado, M. R., & Phelps, E. A. (2011).* How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences of the United States of America, 108*(1), 1–6. http://doi.org/10.1073/pnas.1014938108

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. http://doi.org/10.1073/pnas.0608842104

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013).† Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406–19415. http://doi.org/10.1523/jneurosci.2334-13.2013

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2012).† The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6). http://doi.org/10.1093/scan/nss040

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655–663. http://doi.org/10.1016/j.neuron.2006.03.040

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

Morelli, S. A., Leong, Y. C., Carlson, R. W., Kullar, M., & Zaki, J. (2018). Neural detection of socially valued community members. *Proceedings of the National Academy of Sciences, 115*(32), 8149–8154.

Morelli, S. A., Ong, D. C., Makati, R., Jackson, M. O., & Zaki, J. (2017). Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*(37), 9843–9847. http://doi.org/10.1073/pnas.1702155114

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), 3412–3419.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14*(6), 769–776. http://doi.org/10.1016/j.conb.2004.10.016

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*(1), 3–11. http://doi.org/10.1093/scan/nsm005

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102. http://doi.org/10.1038/nn1968

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications, 9*(1), 332.

Parkinson, C., Liu, S., & Wheatley, T. (2014).† A common cortical metric for spatial, temporal, and social distance. *Journal of Neuroscience, 34*(5), 1979–1987. http://doi.org/10.1523/jneurosci.2159-13.2014

Phan, K. L., Sripada, C. S., Angstadt, M., & McCabe, K. (2010).† Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences of the United States of America, 107*(29), 13099–13104. http://doi.org/10.1073/pnas.1008137107

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron, 48*(2), 175–187. http://doi.org/10.1016/j.neuron.2005.09.025

Rodman, A. M., Powers, K. E., & Somerville, L. H. (2017). Development of self-protective biases in response to social evaluative feedback. *Proceedings of the National Academy of Sciences of the United States of America, 114*(50), 13158–13163. http://doi.org/10.1073/pnas.1712398114

Schiffer, A.-M., Siletti, K., Waszak, F., & Yeung, N. (2017). Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *NeuroImage, 146*, 626–641. http://doi.org/10.1016/j.neuroimage.2016.08.057

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009).† A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508–514. http://doi.org/10.1038/nn.2278

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*(5), 264–268.

Sip, K. E., Smith, D. V., Porcelli, A. J., Kar, K., & Delgado, M. R. (2015). Social closeness and feedback modulate susceptibility to the framing effect. *Social Neuroscience, 10*(1), 35–45. http://doi.org/10.1080/17470919.2014.944316

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006).† Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience, 9*(8), 1007–1008. http://doi.org/10.1038/nn1728

Spunt, R. P., & Adolphs, R. (2017). The neuroscience of understanding the emotions of others. *Neuroscience Letters, 693*, 1–5. http://doi.org/10.1016/j.neulet.2017.06.018

Stanley, D. A. (2016).† Getting to know you: General and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience, 11*(4), 525–536. http://doi.org/10.1093/scan/nsv145

Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron, 80*(3), 816–826. http://doi.org/10.1016/j.neuron.2013.10.038

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7710–7715. http://doi.org/10.1073/pnas.1014345108

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015).† Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences of the United States of America, 112*(5), 201414715. http://doi.org/10.1073/pnas.1414715112

Sul, S., Lu, Güroğlu, B., Crone, E. A., & Chang, L. J. (2017). Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. *Scientific Reports, 7*(8510), 1–10. http://doi.org/10.1038/s41598-017-08692-6

Sutton, R., & Barto, A. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Suzuki, S., Jensen, E. L. S., 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 Sciences of the United States of America, 113*(14), 3755–3760. http://doi.org/10.1073/pnas.1600092113

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016).† Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 194–199. http://doi.org/10.1073/pnas.1511905112

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences of the United States of America, 114*(23), 5982–5987. http://doi.org/10.1073/pnas.1616056114

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008).† Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3*(2), 119–127. http://doi.org/10.1093/scan/nsn009

Uchino, B. N. (2009). Understanding the links between social support and physical health. *Perspectives on Psychological Science, 4*(3), 236–255.

Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598. http://doi.org/10.1111/j.1467-9280.2006.01750.x

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience, 33*(3), 1099–1108.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7*(6), 464–476. http://doi.org/10.1038/nrn1919

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

Zerubavel, N., Bearman, P. S., Weber, J., & Ochsner, K. N. (2015).† Neural mechanisms tracking popularity in real-world social networks. *Proceedings of the National Academy of Sciences, 112*(49), 15072–15077. http://doi.org/10.1073/pnas.1511477112

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage, 129*(C), 95–104. http://doi.org/10.1016/j.neuroimage.2016.01.040